NEW AB INITIO METHODS OF SMALL GENOME SEQUENCE INTERPRETATION

A Thesis
Presented to
The Academic Faculty

By

Ryan E. Mills

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics

Georgia Institute of Technology

May, 2006

NEW AB INITIO METHODS OF SMALL GENOME SEQUENCE INTERPRETATION

Approved by:

Dr. Mark Borodovsky, Advisor
School of Biology
Deparment of Biomedical Engineering
Georgia Institute of Technology

Dr. Jung Choi
School of Biology
Georgia Institute of Technology

Dr. Eberhard Voit
Department of Biomedical Engineering
Georgia Institute of Technology

Dr. Eva Lee
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Allan Tannenbaum
Department of Biomedical Engineering
Georgia Institute of Technology

Date Approved: April 06, 2006

ACKNOWLEDGEMENTS

I would like to acknowledge the guidance of my advisor,
Dr. Mark Borodovsky, and the patience of my wife, Kristin.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The past decade has seen many advances in biological sciences. Complete genetic sequences have been determined from organisms as simple as bacteria and as complex as human. The elucidation of these underlying blueprints of life sets the foundation for further exploration into the inner mechanisms of the cell. However, genomic sequences of free-living organisms alone do not cover the whole spectrum of genomic material.

Viruses are small, non free-living species which require a host to activate and multiply. They range from smaller than 1,000nt in satellite viruses to over 800,000nt in Mimivirus, the largest known virus to date, and are responsible for many of the diseases known in today's society, including influenza and AIDS. Thus, an understanding of the genetic structure of viruses is necessary to find cures for many diseases threatening human health.

This thesis presents novel methods for analysis of short viral sequences and identifying biologically significant regions based on their statistical properties. The first section of this thesis describes the *ab initio* method for identifying genes in viral genomes of varying type, shape and size. This method uses statistical models of the viral protein-coding and non-coding regions. Models are also derived for RBS sequences (for prokaryotic viruses) and the Kozak gene start context (for eukaryotic viruses). We have created an interactive database summarizing the results of the application of this method to viral genomes currently available in GenBank. This database, called VIOLIN, provides an access to the genes identified for each viral genome, allows for further analysis of

these gene sequences and the translated proteins, and displays graphically the distribution of protein-coding potential in a viral genome.

The next two sections of this thesis describe individual projects for two specific viral genomes analyzed with the new method. The first project was devoted to the recently sequenced Herpes B virus from Rhesus macaque. This genome was initially thought to lack an ortholog of the gamma-34.5 gene encoding for a neurovirulence factor necessary for viability of the two close relatives, human herpes simplex viruses 1 and 2. The genome of Rhesus macaque Herpes B virus was annotated using the new gene finding procedure and an in-depth analysis was conducted to find a gamma-34.5 ortholog using a variety of tools for a similarity search. A profound similarity in codon usage between B virus and its host was also identified, despite the large difference in their GC contents (74% and 51%, respectively).

The last thesis section describes the analysis of the Mouse Cytomegalovirus (MCMV) genome by the combination of methods such as sequence segmentation, gene finding and protein identification by mass spectrometry. The MCMV genome is a challenging subject for statistical sequence analysis due to the heterogeneity of its protein coding regions. Therefore the MCMV genome was segmented based on its nucleotide composition and then each segment was considered independently. A thorough analysis was conducted to identify previously unnoticed genes, incorrectly annotated genes and potential sequence errors causing frameshifts. All the findings were then corroborated by the mass spectrometry analysis.

CHAPTER 1

INTRODUCTION

Background

The end of the last century brought about the first foray into the sequencing of complete genomes of free-living organisms. In 1995, the complete genomic sequence of the bacteria *Haemophilus influenzae* was determined (Fleischmann *et al.*, 1995), and many more genomes soon followed (Fraser *et al.*, 1995; Bult *et al*., 1996; Blattner *et al.*, 1997; Tomb *et al.*, 1997; Cole *et al,* 1998). By 2005, the genetic makeups of more complex eukaryotic organisms were elucidated, including the baker's yeast *Saccharomyces cerevisiae* (Cherry *et al.*, 1997), *Caenorhabditis elegans*, (C. elegans Sequencing Consortium, 1998), *Drosophila melanogaster* (Kornberg and Krasnow, 2000), *Homo sapiens* (Lander *et al.*, 2001; Venter *et al.*, 2001), *Mus musculus* (Waterston *et al.*, 2002), and *Pan troglodytes* (Chimpanzee Sequence Consortium, 2005).

The common names of these organisms are by and large recognizable by the general public and thus their genome sequencing was received with a significant amount of media attention. However, this progress was in fact preceeded by the genomic sequencing of smaller, non free-living organisms. In 1977, the complete DNA sequence of the genome of bacteriophage phi X174 was determined, the first viral genome (Sanger *et al.*, 1977). This sequence is made up of 5,375 nucleotides, and helped researchers gain insight into the features responsible for the behavoir of this and other viruses. Since 1997, genomes of many more viruses, prokaryotic and eukaryotic, have been sequenced by various techniques.

These were magnificent achievements given their scale and the technology available at the time; however, the sequences alone are uninformative. The next step is to complete the tedious but monumental task of uncovering underlying functional organization of genomic data, which begins with the problem of gene identification.

A number of gene finding methods have been invented for this purpose. These methods make use of the notion that the ordering of nucleotides in DNA (or, alternatively, of amino acids in proteins) can be described using mathematical means. This ordering in genetic structures such as genes and regulatory regions has statistical patterns which can be determined, modeled, and compared.

There are two major approaches to gene identification, intrinsic and extrinsic (Borodovsky *et al.*, 1994). The intrinsic approach, also called an *ab initio* statistical approach, uses statistical patterns of nucleotide frequencies and nucleotide ordering observed in a given genome. These patterns are not the same in protein-coding and non-coding DNA sequences of a given genome; hence a properly trained intrinsic method can recognize protein-coding regions. Many *ab initio* gene finding tools use DNA information only (Shepherd, 1981; Fickett 1982; Staden 1984; Gribskov *et al,* 1984; Borodovsky *et al*, 1986a; Fields & Soderlund, 1990, Gelfand 1990; Uberbacher & Mural, 1991; Guigo *et al*., 1992; Hutchinson & Hyden, 1992; Borodovsky & McIninch, 1993; Gelfand & Roytberg 1993; Milanesi *et al*., 1993; Snyder & Stormo 1995, 1995b; Dong & Searls, 1994; Soloviev *et al*., 1994; Thomas & Skolnick, 1994; Krogh *et al*., 1994a; Xu *et al*., 1994; Kulp *et al*., 1996; Burge & Karlin, 1997; Krogh 1997; Henderson *et al*., 1997; Zhang, 1997; Lukashin & Borodovsky, 1998; Salzberg *et al*., 1998; Delcher *et al.,* 1999; Shamtkov *et al*., 1999; Reese *et al.* 2000; Salamov and Solovyev, 2000; Parra et al.,

2000; Majoros *et al.*, 2003; Stanke and Waack, 2003). Hidden Markov models (HMM) form the core of several algorithms and have been quite successful. HMMs for gene finding were introduced by Krogh *et al.* (1994). This powerful approach was further developed by Kulp *et al.*, 1996; Burge & Karlin, 1997; Krogh 1997; Lukashin & Borodovsky, 1998.

Other important gene finding methods consider making use of additional, extrinsic information. These extrinsic methods seek to identify evolutionarily conserved sequences in protein-coding regions. These sequences can be detected by similarity searches, mainly against protein databases. The extrinsic method is thus dependent on external information residing outside the sequence of interest (Gish & States, 1993; Borodovsky *et al.*, 1994, Robison *et al.*, 1994; States & Gish, 1994; Burset & Guigo, 1996; Claverie, 1996; Gelfand *et al.*, 1996; Rogozin *et al.*, 1996; Kulp *et al.,* 1997; Frishman *et al.*, 1998; Mironov *et al.*, 1998; Batzoglou *et al.*, 2000). The extrinsic type algorithms have been gaining more attention as the wealth of the extrinsic information grows (Batzoglou *et al.*, 2000; Birney and Durbin, 1997, 2000; Gotoh, 2000; Usaka and Brendel, 2000). There is a special place for the algorithms that align a DNA sequence to a cDNA sequence such as SIM4 (Florea *et al.*, 1998), GeneSequer (Usaka *et al.*, 2000), AAT (Huang *et al.*, 1997) and Spidey (Wheelan *et al.*, 2001), see also Volfovsky *et al.*, 2003; GeneSeqer@PlantGDB (Schlueter et al., 2003); EbEST (Jiang and Jacob, 1998), EST_GENOME (Mott, 1997) and TAP (Kan *et al.*, 2001). The higher degree of conservation of protein-coding regions makes it possible to identify genes in a comparative manner, just by the detection of conservation in syntenic regions. Several

algorithms that use genomic alignments to predict gene structure have also been developed recently (Cawley *et al.*, 2003; Majoros *et al.*, 2005).

Intrinsic and extrinsic methods have complementary strengths. Tests of their predictive power performed on the sets of sequences containing known genes show that the intrinsic methods have higher sensitivity than the extrinsic methods which usually have higher specificity. Using intrinsic and extrinsic methods in concert is therefore a worthwhile approach (Borodovsky *et al.*, 1994).

An improved method for viral gene annotation

Currently, the complete genome of a virus can be sequenced within days. The next step towards the general goal of understanding the details of a virus life cycle is to identify the whole complement of viral genes and proteins. This information can provide critical insights for some other occasions as well. For instance, for a team working on an antiviral drug design, promising drug targets would be those viral proteins that are basically identical in all major strains of a virus and are significantly different from the proteins in the host, e.g. human.

So far, the use of computational gene identification methods in viral genomes by the groups of researchers submitting genomic data to GenBank was primarily restricted to similarity searches. To reduce the risk of missing real genes, a simple rule taking into account the difference in length distributions of real genes and random open-reading frames (ORFs) is frequently applied. This rule suggests annotating ‘long‘ ORFs as genes. For instance, in the rat cytomegalovirus genome any ORF longer than 300 nt not overlapping an adjacent ORF to an extent larger than 60% was annotated as a gene (Vink

et al., 2000). Such a simple rule, however, could cause substantial over-annotation, especially in genomes with high G+C content.

Another frequently used simplification is the annotation of a gene start by the 'longest ORF' rule (assignment of a gene start to the 5'-most ATG codon). A screening of GenBank identified 26 complete viral genomes with a total of 4400 genes, all annotated using this rule. It was nevertheless argued earlier that in a genome with unbiased composition the true start may not be pinpointed by this rule in 25% of cases (Besemer *et al.*, 2001).

Viral genomes are different from the genomes of their hosts in several aspects that hamper immediate successful application of the gene finding methods developed for their hosts. An important factor is the rather small size of a viral genomic sequence. A shorter genome size makes it either impossible to apply previously developed training procedures to derive parameters of the high order statistical models (for the shortest viral genomes) or significantly limits the precision of estimation of parameters of these models (even in the case of the longest viral genomes). Another important feature of the viral genome organization is the high frequency of gene overlaps that occur in viruses of both prokaryotic and eukaryotic hosts. The gene overlaps in viral genomes appear to be considerably longer than those seen in prokaryotic and, much more rarely, eukaryotic genomes. Furthermore, some annotated and experimentally confirmed viral genes may completely overlap each other. Repetitive DNA may occupy a large portion of a viral genome; for example, in the Epstein–Barr virus genome (NC_001345 [GenBank]) repetitive regions amount to 30% of the genomic sequence (Kieff and Rickonson, 2001),

thus making the size of the sequence available for model training shorter than could be expected.

In spite of the difficulties mentioned above, several genome sequencing groups have attempted to apply earlier developed statistical gene prediction programs for viral genome annotation. For instance, the GeneMark program (Borodovsky and McIninch, 1993a) was used to identify genes in the genomes of Bovine herpesvirus 4 (Zimmerman *et al.*, 2001), bacteriophage FKZ of Pseudomonas aeruginosa (Mesyanzhinov *et al.*, 2002), Mycoplasma virus P1 (Tu *et al.*, 2001), Mycobacteriophage D29 (Ford *et al.*, 1998), Stx 2e-encoding phage FP27 (Recketenwald and Schmidt, 2002), coliphage T4 and the marine cyanophage S-PM2 (Hambley *et al.*, 2001), as well as to identify genes in genomes of virulence plasmids in Rhodococcus equi (Takai *et al.*, 2000), Shigella flexneri (Venkatesan *et al.*, 2001) and Escherichia coli (Burland *et al.*, 1998). Still, these initial attempts did not use a tool developed specifically for the problem in hand (except perhaps the case of T4, where the GeneMark models were adjusted to the genomic T4 sequence).

A significant difference may exist sometimes between the GenBank record and the original publication. For instance, the annotation of the white spot bacilliform virus (AF332093 [GenBank]) lists 531 protein-coding genes in comparison with only 181 genes mentioned in the original publication (Yang *et al.*, 2001). On the other hand, only 23 genes are annotated in Rana tigrina ranavirus (AF389451 [GenBank]), while the original publication (He *et al.*, 2002) describes 105 genes. In order to improve the quality of DNA sequence annotation, the National Center for Biotechnology Information (NCBI) has created the RefSeq collection. While the original GenBank genomic record is

maintained as suggested by the authors, the RefSeq record of the same sequence is continuously updated with regard to new relevant data that become available. There were 1766 RefSeq records for complete genomes of viruses of prokaryotic and eukaryotic hosts as of March 2006.

Several attempts have been made to organize data on viral genomes in interactive databases providing tools for analysis of viral genes and proteins (Farmer *et al.*, 1995; Hiscock and Upton, 2000; Mar Alba *et al.*, 2001). These projects have been typically focused on specific classes of viruses.

Gene annotation in viruses often relies upon similarity search methods. These methods possess high specificity but some genes may be missed, either those unique to a particular genome or those highly divergent from known homologs. To identify potentially missing viral genes we have analyzed all complete viral genomes currently available in GenBank with the augmented version of gene finding program GeneMarkS. In particular, by implementing genome-specific self-training protocols we have better adjusted the GeneMarkS statistical models to sequences of viral genomes. Viral genome properties depend on their class; for instance, ssRNA positive strand viruses have genes only in one strand. Other considerations include whether the virus is linear or circular. Therefore, the self-training algorithm, GeneMarkS, was changed to take these aspects into account. In addition, modifications were also made in the gene finding algorithm itself to allow for using the Kozak gene start context model for eukaryotic viruses. This gene start context model, and the corresponding RBS used by the original GeneMarkS, could be derived by the training procedure itself. The new procedure was used for an analysis of more than 5000 viral sequences obtained from GenBank. The VIOLIN

database of predicted genes was compiled. This database provides access to the gene predictions and their comparison with the current annotation. The database includes capabilities for running domain and homology searches directly from the webpage. This database and its recent updates (Mills et al., 2003) are described in Chapter 2.

Experimental and computational approaches for analyzing viral genomes

Just as intrinsic and extrinsic methods can be used in concert to improve the prediction of genes, so experimental and computational methods can be used together to analyze the genomic structure and the protein complement of a virus. Chapters 3 and 4 present the projects involving the sequencing and analysis of the *Macaca mulatto* B virus and the analysis of *Mus musculus* cytomegalovirus virions. These projects were done in collaboration with experimental laboratories; therefore the experimental studies were conducted along with the computational analyses.

Sequencing and Analysis of the Rhesus macaque B virus

The study of the complete genome of herpes B virus (Perelygina et al, 2003) uses the methods we developed within the VIOLIN project (Chapter 3). Comparative genome analyses of closely related viruses offer insights into the conserved patterns of gene distribution in viral genomes (Dolan *et al.*, 1998; Virgin *et al.,* 1997), the phylogenetic relationships (Bair and Darai, 2001; Dominguez *et al.*, 1999) and into evolution of viral genes involved in virulence and pathogenicity (Afonso *et al.*, 2001; Kingham *et al.*, 2001; Tulman *et al.*, 2000). Currently there are 42 completely sequenced herpesvirus genomes (GenBank data, http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html).

The B virus (Cercopithecine herpesvirus 1, monkey B virus) is a member of the subfamily Alphaherpesvirinae, which together with human herpes simplex virus types 1 and 2 (HSV-1 and HSV-2) constitutes the genus Simplexvirus. The B virus generally causes only mild localized or asymptomatic infections in its natural hosts, Asian monkeys of the genus Macaca (Keeble, 1960; Keeble *et al.*, 1958; Whitley and Hilliard, 2001). In contrast, B virus infections in foreign hosts, humans or monkey species other than macaques, often result in encephalitis, encephalomyelitis, and death (Palmer, 1987; Weigler, 1992; Whitley and Hilliard, 2001).

The genome organization of the B virus is similar to that of HSV-1 and HSV-2: the unique long (UL) and unique short (US) segments flanked by inverted long (RL) and short (RS) repeat sequences are covalently joined in four possible isomeric configurations (Harrrington *et al.*, 1992). Sequence analysis of the partial US regions of the B virus and simian agent 8 virus demonstrated that human and primate viruses are colinear in this genomic segment (Ohsawa *et al.*, 2002). However, only a limited number of the B virus gene sequences from the UL region have been published (Bennett *et al.*, 1992; Killeen *et al.*, 1992; Slomka *et al.*, 1995; Smith *et al.*, 1998), and nothing has been reported about the structure and the gene content of the repeat genomic elements.

The size of B virus genome was estimated previously as 165 kb (Harrrington *et al.*, 1992) and 162.5 kb (Ludwig *et al.*, 1983), which is significantly larger than the HSV-1 and HSV-2 genomes (152 and 155 kb, respectively). The extra DNA might contain additional genes that are not present in human viruses and may provide insight about the B virus pathogenicity in foreign hosts.

Analysis of Mus musculus cytomegalovirus virions

Computational methods of viral genome analysis can be efficiently complemented by new high-throughput methods such as mass spectrometry as described in Chapter 4. Particularly, proteins associated with the MCMV viral particle were identified by a combined approach using genomic and proteomic methods (Kattenhorn et al, 2003).

Murine cytomegalovirus (MCMV), a member of the betaherpesvirus family, shares 45.2% sequence identity with human CMV (HCMV) and is currently the most commonly used model for the study of CMV-induced disease. The MCMV genome has a size of 230 kbp and was originally estimated to encode 170 proteins (Rawlinson *et al.*, 1996). The MCMV virion consists of double-stranded viral DNA surrounded by an icosahedral capsid, a complex proteinaceous tegument, and a lipid membrane (Mocarski and Courcelle, 2001). Although the composition of HCMV particles has been studied (Baldick and Shenk, 1996; Bresnahan and Shenk, 2000), the protein composition of the MCMV virion has not been known in detail.

The majority of MCMV genes have been annotated computationally using homologues in HCMV (Rawlinson *et al.*, 1996). Consequently, many open reading frames (ORFs), including genes encoding some structural proteins, lack experimental confirmation. Analysis of regions unique to MCMV has been even more limited. There has been, therefore, a need in confirmation of putative structural and functional homologues of HCMV gene products as well as in characterization of the composition of MCMV particles.

CHAPTER 2

IMPROVING GENE ANNOTATION IN COMPLETE VIRAL GENOMES

Introduction

Gene annotation in viral genomes often relies upon similarity search methods. These methods possess high specificity but some genes may be missed, either those unique to a particular genome or those highly divergent from known homologs. To identify potentially missing viral genes we have analyzed all complete viral genomes currently available in GenBank with a specialized and augmented version of the gene finding program GeneMarkS. In particular, by implementing genome-specific self-training protocols we have better adjusted the GeneMarkS statistical models to sequences of viral genomes. Hundreds of new genes were identified, some in well studied viral genomes. For example, a new gene predicted in the Epstein-Barr virus genome was shown to encode a protein similar to a-herpesvirus minor tegument protein UL14 with heat shock functions. Convincing evidence of this similarity was obtained after 12 PSI-BLAST iterations. In another example, several iterations of PSI-BLAST were required to demonstrate that a gene predicted in the genome of Alcelaphine herpesvirus 1 encodes a BALF1-like protein which is thought to be involved in apoptosis regulation and, potentially, carcinogenesis. New predictions were used to refine annotations of viral genomes in the RefSeq collection curated by the National Center for Biotechnology Information. Importantly, even in those cases where no sequence similarities were detected, GeneMarkS significantly reduced the number of primary targets for experimental characterization by identifying the most probable candidate genes.

As a result of the application of this tool, we have created new annotation records for viral genomes present in GenBank. These records have been compiled in the database VIOLIN (Mills *et al.*, 2003). The database interface allows for retrieving the results of comparative analysis of the predicted genes as well as viewing a graphical representation of the protein-coding potential of the genomic regions around each predicted gene as well as of the full genome. Access to individual nucleotide and protein prediction sequences is also available directly from the database interface. VIOLIN is accessible online at http://opal.biology.gatech.edu/GeneMark/VIOLIN/.

Methodology

For phage genomes with prokaryotic-type gene organization, computer methods of prokaryotic gene finding could be adjusted rather easily. The prokaryotic version of GeneMark.hmm (GeneMark.hmm-P) as well as its self-training version GeneMarkS (GeneMark.hmm-PS) were previously shown to possess high accuracy both in detecting prokaryotic genes as a whole and in exactly pinpointing gene starts (Lukashin and Borodovsky, 1998; Besemer and Boroodvsky, 1999). Therefore, GeneMarkS was the natural choice as the tool to be adjusted for the analysis of phage genomes. For viruses of eukaryotic hosts, the situation is more complex. Current eukaryotic gene finding algorithms are unable to predict the gene overlaps frequently seen in genomes of viruses of eukaryotic hosts. On the other hand, according to the RefSeq annotation of 23,474 genes in 1545 genomes of viruses of eukaryotic hosts, only 421 genes have introns. Therefore, use of the program able to predict overlapping genes provides more benefits than the one predicting exon–intron structures. The program suitable with necessary

modification and subsequent use was again the prokaryotic GeneMarkS, which could identify overlapping protein-coding ORFs while rarely occurring exons would be predicted as separate ORFs.

A viral genomic sequence might not provide enough training data to determine parameters of Markov chain models used in GeneMark.hmm. We turned, therefore, to the heuristic training technique (Besemer and Borodovsky, 1999), which is able to derive the parameters of the required models from a DNA sequence as short as 400 nt.

For large viral genomes, the statistical models initially defined by the heuristic procedure could be iteratively refined further by the unsupervised training procedure implemented in GeneMarkS (Besemer *et al.*, 2001). This iterative procedure uses simultaneous training and gene prediction to build models of protein-coding and non-coding sequences. For large phage genomes, GeneMarkS derived a model for the ribosomal binding site (RBS) and its spacer (the sequence between the rightmost nucleotide of the RBS and the first nucleotide of the start codon). Parameters of both models were determined from the multiple alignment of the nucleotide sequences situated upstream of the predicted gene starts, with the alignment constructed by the Gibbs Motif Sampler (Lawrence *et al.*, 1993). For large enough genomes of viruses of eukaryotic hosts, parameters of a model for the Kozak pattern associated with the translational initiation site were determined by GeneMarkS with yet another modification. This GeneMarkS version allowed for using the Kozak model for gene start prediction. Further modifications were done to adjust the program to different types of viral genome organization.

Since a linear viral genome cannot have a partial coding region at either terminus, a specific restriction imposed at the program initialization stage excluded this possibility. Conversely, an additional post-processing step was implemented for circular viral genomes to detect genes possibly divided by the split point chosen in the original annotation. For the single-stranded RNA (ssRNA) positive strand viruses whose genes are located in one strand only, an additional procedure identified the strand where gene predictions clustered predominantly and the opposing strand was assigned as completely non-coding.

For each viral genome the training procedure had to determine whether the sequence data were sufficient for obtaining heuristic a full training cycle of GeneMarkS. If GeneMark.hmm with the initially defined heuristic models predicted fewer number of genes than the threshold number Nm, then the procedure stopped and these initial predictions were not refined further. Otherwise, the full cycle of GeneMarkS training was initiated. The default Nm number was defined as 50.

In the training process, if several repetitive copies of some predicted protein-coding ORFs were identified, all copies but one were excluded from the training set of protein-coding regions to reduce bias in the protein-coding sequence model. Predicted ORFs longer than 500 nt that appeared in predicted intergenic regions were removed from the set of non-coding regions to exclude possible 'contamination' of the non-coding training set. For viral genomes with a total size of predicted non-coding regions <10 kb, the training set of non-coding regions was augmented with an additional 10 kb sequence generated by the simple multinomial model (Durbin *et al.*, 1998), simulating a sequence

with the frequencies of the four nucleotides identical to those observed in the non-coding region of the genome.

The step-wise diagram of the GeneMarkS self-training and gene prediction for a phage genome is shown in Figure 2.1. Note that for a virus of a eukaryotic host, the Kozak context model is replacing the RBS model.

The evaluation of the RBS model quality was done by assessing both the variance of the RBS signal localization and the information content of the RBS model derived by the Gibbs Sampler. The Kozak context model was evaluated in a similar manner.

The self-training procedure was terminated as soon as two subsequent iterations produced the same set of gene predictions. However, in some cases exact convergence was not achieved due to small cyclic variations observed in subsequent iterations. In this case the self-training process was stopped and the sequence parse into coding and non-coding regions was taken to be the one with the larger number of predicted genes.

The BLAST searches used to characterize newly predicted proteins were conducted using default parameters: BLOSUM62; penalty for gap ‘10’; penalty for gap extension ‘1’; low-complexity filtering ‘on’. In PSI-BLAST searches, the parameters were the same with the exception that the low-complexity filtering was ‘off’.

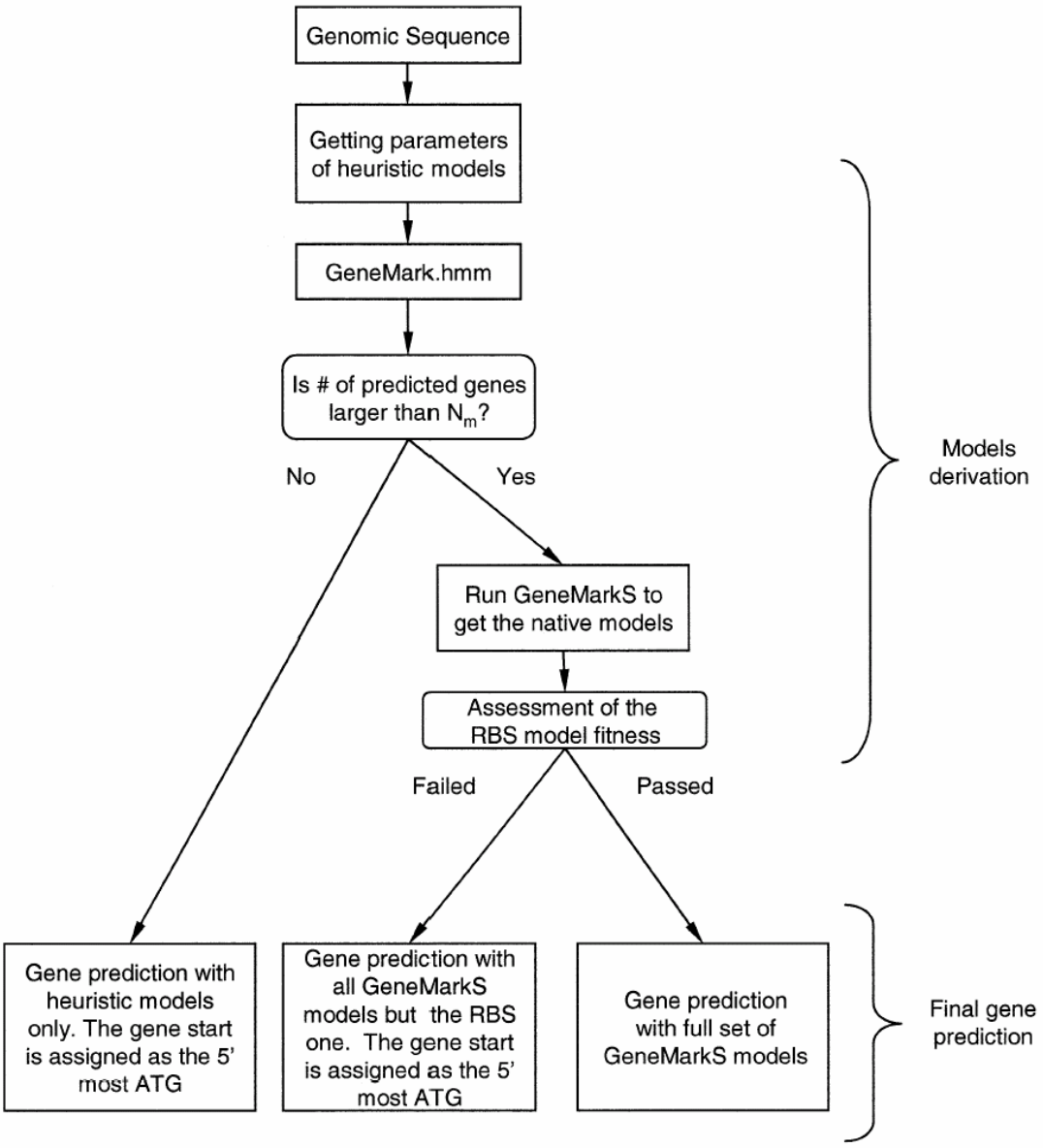

Figure 2.1 - Flowchart of the statistical gene identification procedure applied to a complete genome of a virus of a prokaryotic host. For viruses of eukaryotic hosts, the Kozak context model is used instead of the RBS model.

Results

The VIOLIN database

The VIOLIN database has increased dramatically since its original inception, growing from approximately 3000 to over 6000 complete genomic records as obtained from GenBank (Benson *et al.*, 2005). This growth is attributed to an increase in the number of completely sequenced genomes to GenBank, as well as the addition of individual strains and isolates of known viruses. We have obtained many of these new genomic records as well as updated records for preexisting genomic annotations in our database from the GenBank repository.

Currently the VIOLIN database includes analyses of 6127 complete viral genomes. Its CGI web frontend is connected to a MySQL backend database. Users can access individual viral genome records by using the NCBI accession number or by using the search based on the viral definition given in the GenBank record (Figure 2.2)

Each viral record contains a feature list displaying the coordinates of predicted genes as well as genes annotated in GenBank but not found by our analyses. A graphical report for each genome and individual gene prediction therein allows for viewing the distribution of protein-coding potential for each locus of the viral genome as defined by GeneMark using models specific from this virus**.** VIOLIN allows for direct accessing nucleotide and protein sequences for each individual gene prediction and to for using the protein sequences in PSI-BLAST and RPS-BLAST analyses. The sequences are provided in FASTA format.

As stated previously for viruses of eukaryotic hosts the GeneMarkS procedure does not attempt to find the structure of intron-containing genes. However, such

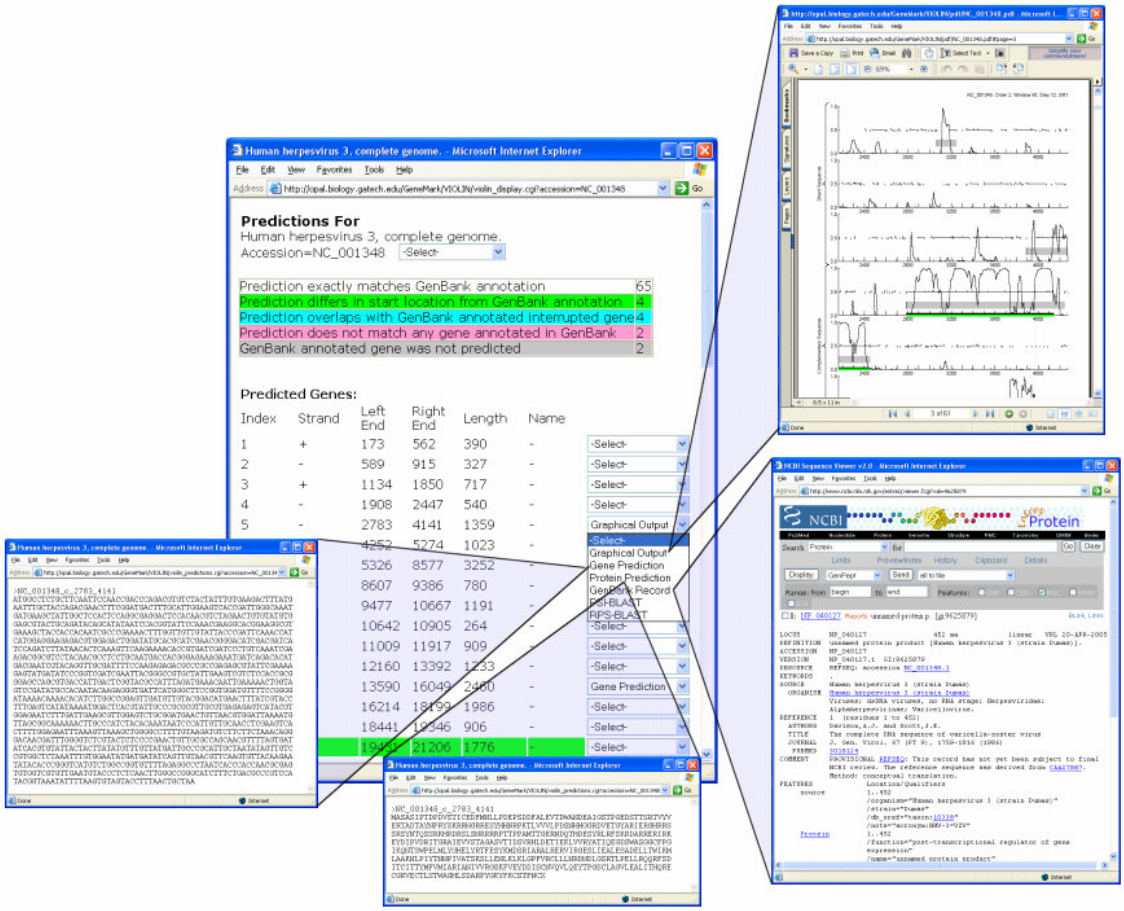

Figure 2.2– Zooming in into the VIOLIN database record

structures still can be inferred graphically by combining GeneMark illustrated predictions. To help facilitate identification of such genes, a section is included in the feature list of VIOLIN which identifies introns containing genes, as annotated in the GenBank record. These genes are labeled on the graphical display and are overlaid with the most current GenBank annotation for comparisons with the predicted genes. The genomic sequence is represented in 3 frames for each strand, with the distribution of protein-coding potential, the *a posteriori* probability of protein coding function, displayed. This graphical representation can be indicative of the presence of unusual features of gene organization.

For many viral species complete genomic sequences were determined for several genomic variants (strains, mutants, isolates). Therefore, the whole set of complete viral genome records contains many almost identical entries. The RefSeq collection normally exemplifying one genomic sequence per viral species presents an arguably non-redundant subset of viral genomes with 1,545 eukaryotic and 221 prokaryotic virus records (phages) as of March 2006. Thus, in what follows, the subset of VIOLIN corresponding to the viral genomes from the RefSeq collection will be our focus. The statistics of the VIOLIN entries were gathered by a series of SQL queries to the MySQL database in which the records are stored.

The overall figures characterizing the gene finding results for the RefSeq viral genomes are as follows. In 1,545 complete genomes of eukaryotic viruses 15,243 predicted protein-coding genes matched the existing annotation exactly. In addition, 623 predicted genes had the same strand, reading frame, and translation stop as an annotated gene, while the positions of the translation start were different. In 475 cases predicted

coding regions overlapped known intron containing genes. Notably, 2,868 gene predictions not matching any previously annotated gene carried entirely new information. Evolutionary conserved regions were detected in 1,060 of these newly predicted proteins by BLASTP search with P-value < 10^{-5}. In 749 out of these 1,060 cases, the newly predicted genes did not have any overlap with annotated genes. 311 other gene predictions contained overlaps with annotated genes but the reading frames were different. A rather large number of 3,422 total genes annotated in the viral genomes of eukaryotic hosts were not confirmed by our analysis.

In 221 phage genomes from the RefSeq collection, 9,351 predicted protein-coding genes matched the annotation exactly. Additionally, 1,780 gene predictions had the same strand, reading frame and translation stop as an annotated gene, but different position of the gene start. Overlaps between predictions and intron-containing genes were observed in 25 cases. 917 other gene predictions carried entirely new information. In 394 of these newly predicted proteins evolutionary conserved regions were detected by BLASTP search with P-value < 10^{-5}. In 200 out of these 394 cases predicted genes did not have any overlap with an annotated gene. Each one of the other 194 predicted genes overlapped with an annotated gene but the reading frames were different. This analysis did not confirm 1,574 annotated phage genes.

Subsequent BLASTP analysis of protein products of the viral genes annotated in 1,766 viral genomes from the RefSeq collection but not predicted produced the following results. No similar protein other than itself was found for protein products of 1,687 genes out of 3,422 genes annotated in eukaryotic viruses and for 1,136 genes out of 1,810 genes annotated in phages. This analysis indicates that the total number of false negative

predictions might be as low as 2409 (1735 + 674) which is less than 10% of the whole number of predicted genes in all viral genomes from the RefSeq collection. Interestingly, no single annotated gene was missed in the analysis of over 620 viral genomes.

The statistics of the BLASTP search results for the genes predicted and annotated in the RefSeq genomes as well as in all other viral genomes in GenBank is shown in Table 2.1. In comparison with the genomes from the RefSeq collection, analysis of the GenBank genomes produced a larger fraction of new genes (20.2% of the total number of predicted genes). Also, among these new genes, those with a significant BLAST hit constitute 9.2% versus the 5.7% fraction characteristic for the new genes predicted in the RefSeq genomes.

The numbers given in Table 2.1 for genomes from the RefSeq collection could be grouped together by virus length and type (Tables 2.2 and 2.3, respectively). Interestingly, a large number of new genes were identified in genomes shorter than 10,000nt, though this is the largest group with 1,344 genomes. For example, in the 8,454nt long genome of the single stranded DNA enterobacteria phage IF1 (NC_001954) a new gene was identified with length 204nt. This gene was encoded in a DNA strand complementary to the one physically present in the viral particle.

The largest numbers of genes that have been confirmed, updated or newly identified, as it is seen from Table 2.3, belong to double stranded DNA viruses (389 genomes) and single stranded RNA viruses (646 genomes).

Assessment of the accuracy of an automatic gene annotation is the critically important issue. Given that deviation of gene predictions from “the truth” can go both sides, either as over prediction or under prediction, there are, traditionally, two

Table 2.1 - Summary statistics of the complete VIOLIN database as of March 2006.

	GenBank Total	RefSeq Total
Database summary		
Number of viral genomes analyzed	6127	1766
Prediction and annotation comparison		
Exact match between prediction and annotation	55987	25980
Predicted gene differs in start location from annotated gene	6039	2533
Predicted gene overlaps with an intron containing annotated gene	1840	502
Annotated gene was not predicted (possible false negative)	13139 (23.4%)[a]	5564 (21.4%)[a]
Newly predicted gene (possible false negative)	11331 (20.2%)[a]	3919 (15.1%)[a]
Analysis of newly predicted genes		
Prediction has a BLASTP hit with *E*-value < 0.005	5202	1504
Prediction has no BLASTP hit with *E*-value < 0.005	6129	2415

[a]The percentage value is defined with regard to the number of predicted genes exactly matching the annotation in GenBank.
[b]The percentage value is defined with regard to the number of predicted genes exactly matching the annotation in RefSeq.

Table 2.2 - Summary statistics of the RefSeq section of VIOLIN by virus size as of March 2006.

b	L < 10000 nt[a] (1344)[c]	10000 nt <= L <= 100000 nt[a] (315)[c]	L > 100000 nt[a] (107)[c]
Exact match	2613	7861	14353
Different Start	309 (11.8%)	1560 (19.8%)	554 (3.9%)
Overlap with Interrupted Gene	112 (4.3%)	98 (6.3%)	294 (2.0%)
Annotated Gene Not Predicted	1078 (41.3%)	1537 (19.5%)	2743 (19.1%)
New Predictions	501 (19.2%)	958 (12.2%)	2343 (16.3%)
Analysis of Newly Predicted Genes			
PSI-BLAST hit	138	434	885
No hits	363	524	1458

[a] The genome length is designated as L.
[b] The meaning of the categories in this column is the same as in the right-most column in Table 2.1

Table 2.3 - Summary statistics of the RefSeq section of VIOLIN by virus type as of March 2006.

b	dsDNA (389)[a]	ssDNA (294)[a]	dsRNA (229)[a]	Positive Strand (517)[a]	Negative Strand (129)[a]	Retroid (85)[a]	Satellite (63)[a]	Other Viruses (40)[a]	Other Phages (20)[a]
Exact match	21391	749	265	904	356	220	26	89	827
Different Start	1932	80	10	148	46	37	2	9	159
Overlap with Interrupted Gene	380	7	4	71	7	34	0	0	0
Annotated Gene Not Predicted	4252	454	28	324	46	57	8	11	78
New Predictions	3186	130	80	104	51	63	22	16	150
Analysis of Newly Predicted Genes									
PSI-BLAST hit	1246	51	10	39	6	21	1	6	77
No hits	1940	79	70	65	45	42	21	10	73

[a]The number in parenthesis designates the number of genomes of a given category.
[b]The meaning of the categories in this column is the same as in the right-most column in Table 2.1

characteristics of accuracy, sensitivity and specificity. The sensitivity and specificity values could be objectively determined for DNA sequences with experimentally verified protein coding regions. The value of sensitivity (Sn) is defined as the ratio of the number of true predictions to the number of genes in a test set. The value of specificity (Sp) is defined as the ratio of the number of true predictions to the total number of predictions made.

This approach to accuracy definition makes a rather clear sense for genes of the prokaryotic type, and it is used throughout this study since the vast majority of viral genes are not interrupted by introns. For a prokaryotic gene, it is assumed that the correct prediction (true prediction) takes place if the 3' end of the predicted gene matches the 3' end of the experimentally verified one.

One more characteristic of prediction quality is the accuracy of prediction of gene translation start. The start prediction accuracy value is defined by the fraction of correct gene start predictions (those matching experimentally verified ones) among predictions that matched the 3' end of the annotated gene.

Ideally, the values of Sn and Sp should be determined for the experimentally verified test sets. However, these data sets may not be readily available. Any given viral genome, except perhaps several of a tiny size, has quite a few genes annotated without experimental evidence. With no consistent experimentally validated set available, a test set, the set of "trustable" genes, has been compiled. For instance, in 9 genomes of human herpesviruses (Table 2.4), genes both annotated and predicted were selected as trustable. Those genes that were either annotated or *ab initio* predicted and, in addition, possessed an "extrinsic" evidence for being a real gene, such as the experimentally derived function

Table 2.4 - Summary statistics of the Human Herpesvirus test set

Virus	Number of Genes Predicted	Number of Genes Annotated	Number of Genes in Test Set	Prediction Sensitivity	Prediction Specificity
HHV-1 (HSV-1)	76	73	75	0.92	0.90
HHV-2 (HSV-2)	77	71	71	0.92	0.84
HHV-3 (VZV)	72	71	71	0.97	0.96
HHV-4 (EBV)	90	94	78	0.89	0.78
HHV-5 (HCMV)	164	198	148	0.84	0.76
HHV-6A	115	121	119	0.87	0.90
HHV-6B	114	91	85	0.95	0.71
HHV-7	109	107	104	0.87	0.83
HHV-8 (KSHV)	96	82	88	0.94	0.86
Total	913	908	839	0.91	0.84

annotation or statistically significant similarity to well characterized genes were also considered as trustable. The average figures obtained in this analysis, Sn = .91 and Sp = .84 appeared to be satisfactory. Note that the Sp value is likely to be underestimated due to excluding several genes from the test set while leaving sequence space for possible erroneous predictions.

A comparison of length distributions of genes newly predicted and genes predicted but not annotated (Figure 2.3) shows that the newly predicted genes have a shorter average length. This observation may indicate a bias in the original expert annotation toward longer ORFs. The longer ORFs are assumed to be more likely to be real genes while ORFs shorter than 300nt are difficult to discriminate from random non-coding ORFs. This trend may lead to over prediction of ORFs longer than 300nt as genes, while some short genes may be missed in the annotation. As could be seen from Figure 2.3, many "long" annotated genes indeed are not confirmed by the computer analysis while quite a few "short" gene predictions are added.

The evaluation of the gene start prediction accuracy is another important issue. As described above, to identify gene start related nucleotide sequence patterns additional models were used, the one for RBS (Shine-Dalgarno) and one for the Kozak context pattern for prokaryotic and eukaryotic viruses respectively. The RBS pattern can be found on a variable distance from the gene start. The frequency distribution of this distance, the spacer length, was yet one more statistical model involved. To give an example, the RBS nucleotide frequency patterns specific for phage T4 and phage λ are shown graphically in the "logo" form (Schneider and Stephens, 1990) in Figure 2.4ab. Note that these patterns are

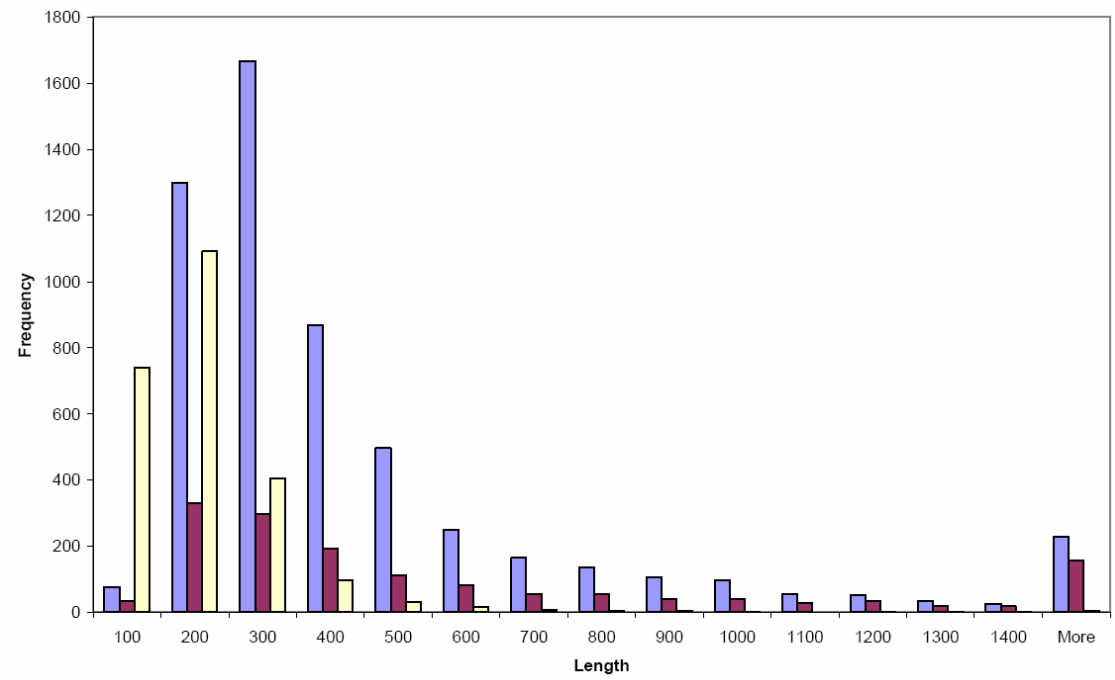

Figure 2.3 - Length distribution of viral genes in the RefSeq section of the VIOLIN database. The genes of the predicted and but annotated category are divided into two groups: with BLASTP hits with P-value < 0.00001 (red) and significant hits (yellow). The genes of the the annotated but not predicted category are in blue.

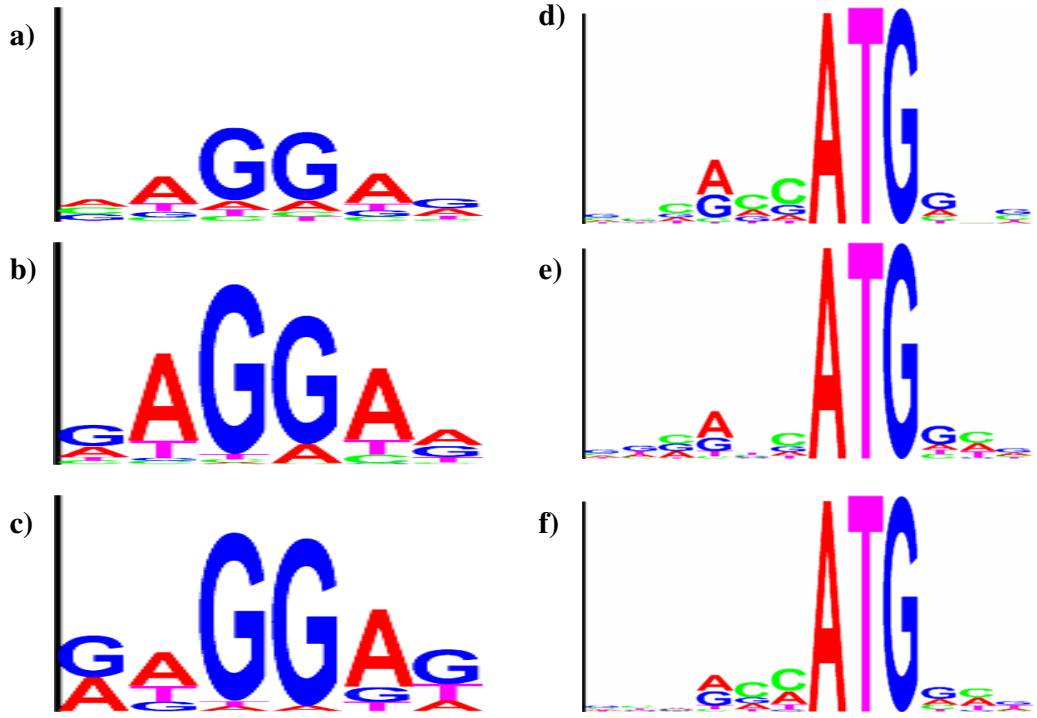

Figure 2.4 - The RBS positional nucleotide frequency patterns of phage T4 and phage λ (panels *b* and *c*, respectively) shown in the logo form (26), can be compared to the RBS pattern of their host *E. coli* (panel *a*). Similarly, the Kozak pattern for human herpesvirus 4 and human herpesvirus 8 (panels *e* and *f*, respectively), can be compared with the Kozak pattern for human genes (panel *c*).

similar to one known for *E. coli* (Figure 2.4c) as would be expected given that these phages are using the *E. coli* translational mechanism. The logos for the Kozak frequency pattern determined for the Epstein-Barr virus (HHV4) and for Karposi's Sarcoma herpesvirus (HHV8) shown in Figure 2.3de indicate that though the information content here is lower than in the RBS signal, the patterns are similar to the Kozak pattern observed in the genome of their human host (Figure 2.4f).

Accurate evaluation of the gene start prediction accuracy requires a set of genes with experimentally verified gene starts. Evaluation of GeneMarkS performance was done earlier on the test set of *E. coli* genes with 5' ends experimentally verified by sequencing of N-terminals of encoded proteins (Link *et al.*, 1997). This test was shown the the accuracy of the gene start prediction is as high as 94% (Besemer and Borodovsky, 2001). No set of N-terminal sequenced proteins is currently available for eukaryotic viruses. Therefore, a set of genes was compiled from nine human herpesviruses with translation starts confirmed by similarity search on a protein level. Self-hits to the same protein annotated in GenBank were ignored. The results of the analysis have shown that due to the variability of the gene and protein sequences in proximity of the gene and protein start the similarity search method do not give reliable indication of the real start. Still, the selection of the most reliable cases gave an estimate of the accuracy of gene start prediction as 85%.

GeneMark.hmm-V has been regularly used by the NCBI curators to improve the annotation of viral genomes in the RefSeq collection of Complete Genomes (Tatusova *et al.*, 1999). Gene predictions have been subjected to additional analysis and by the NCBI staff for quality control and functional assignment. Currently, predictions listed in

VIOLIN were incorporated into annotations of 84 viral genomes in the RefSeq collection. For example, in Fowl adenovirus D (NC_000899) 14 proteins have been added to 15 existing in the original GenBank record AF083975. This was a particularly difficult case because many of the newly added genes were disrupted by frameshifts that likely resulted from sequencing errors. The new tentative protein sequences were assembled from fragments predicted by GeneMarkS using the ORF Finder (Tatusov and Tatusova, unpublished), and BLASTP searches. In another example, lymphocystis disease virus (NC_001824) 110 coding regions were identified while the original GenBank record (AF083975) contained only one gene for major capsid protein. Table 2.5 shows the list of updated gene annotations in RefSeq incorporating the results of our analysis. The VIOLIN project will continue to evolve as more complete viral genome sequences enter public domain and undergo annotation refinement.

Table 2.5 - VIOLIN predictions already incorporated in RefSeq annotations

Group	Prediction	Predicted Length	Best BLASTP hit	BLASTP length	Score	E-value	Annotated Function
dsDNA	**Alcelaphine herpesvirus 1, complete genome.**		**NC_002531**				
	10443..11138	231	gi|9628007|	183	66.3	4.00E-10	Putative BALF1 homolog
	Amsacta moorei entomopoxvirus, complete genome.		**NC_002520**				
	complement(114621..114773)	50	gi|9629968|	52	65.6	9.00E-11	conotoxin-like protein
	Ateline herpesvirus 3, complete genome.		**NC_001987**				
	73911..75053	380	gi|331012|	384	603	e-171	immediate-early phosphoprotein (transactivator)
	Avian adenovirus CELO, complete genome.		**NC_001720**				
	26793..27119	108	gi|9633186|	302	95.6	2.00E-19	Late 33 kDa protein
	Bovine adenovirus 2, complete genome.		**NC_002513**				
	10583..12295	570	gi|13487865|	573	755	0	peripentonal hexon-associated protein
	12347..13783	478	gi|13487866|	471	793	0	penton protein
	15888..16382	164	gi|13487870|	233	201	4.00E-51	minor capsid protein VI precursor
	16628..19324	898	gi|13487871|	910	1546	0	hexon protein
	21366..23579	737	gi|13487873|	722	1004	0	hexon assembly-associated 100 kDa protein
	complement(30406..30735)	109	gi|13487881|	245	101	3.00E-21	245R protein homolog
	complement(30823..31383)	186	gi|13487880|	253	188	5.00E-47	253R protein homolog
	Deer papillomavirus, complete genome.		**NC_001523**				
	3914..4048	44	gi|137747|	44	85.4	9.00E-17	E5 transforming protein
	Equine herpesvirus 1, complete genome.		**NC_001491**				
	complement(112994..113785)	263	gi|15235673|	608	179	5.00E-44	glycine-rich protein
	Fowl adenovirus 8, complete genome.		**NC_000899**				
	14583..16211	542	gi|9628848|	575	799	0	peripentonal hexon associated protein
	complement(38665..40446)	593	gi|3845680|	195	381	e-104	glycine-rich protien
	Fowlpox virus, complete genome.		**NC_002188**				
	52914..54572	552	gi|1083970|	552	1122	0.00E+00	Rifampicin resistance N3L protein
	Human adenovirus type 2, complete genome.		**NC_001405**				
	30444..30830	128	gi|119063|	128	264	5.00E-70	early E3B protein
	complement(30852..31019)	55	gi|9626584|	53	143	4.00E-09	U protein
	complement(35146..35532)	128	gi|119716|	283	246	1.00E-64	E4 protein
	Human adenovirus type 12, complete genome.		**NC_001460**				
	25202..25558	118	gi|9626562|	211	135	2.00E-31	33 kDa phosphoprotein
	complement(31183..31407)	74	gi|93525|	74	154	1.00E-37	early E4 17 kDa protein
	Human adenovirus type 17, complete genome.		**NC_002067**				
	560..1138	192	gi|4323354|	251	316	1.00E-85	Early E1A protein
	1491..2117	208	gi|4323357|	182	377	e-104	small T-antigen fragment
	2165..2533	122	gi|4323358|	495	214	5.00E-55	small T-antigen fragment
	2530..2976	148	gi|4323358|	495	301	5.00E-81	small T-antigen fragment
	3033..3359	108	gi|4323358|	495	227	4.00E-59	small T-antigen fragment
	complement(3888..4499)	203	gi|130244|	448	408	e-113	IVa2 maturation protein
	complement(4501..4935)	144	gi|130244|	448	250	8.00E-66	IVa2 maturation protein
	15724..15960	78	gi|9626191|	368	74.5	2.00E-13	V minor core protein
	16177..16713	178	gi|9626570|	358	148	4.00E-35	V minor core protein
	16798..16953	51	gi|9626571|	70	74.5	2.00E-13	L2 protein mu precursor
	17754..18065	103	gi|780528|	947	161	3.00E-39	hexon capsid protein
	18068..20617	849	gi|780528|	947	1595	0	hexon capsid protein
	complement(21293..21745)	150	gi|118737|	517	238	3.00E-62	E2A DNA binding protein
	complement(21724..22503)	259	gi|118735|	512	341	6.00E-93	E2A DNA binding protein
	23513..23779	88	gi|209871|	652	99.8	7.00E-21	hexon assembly-associated protein
	23799..24956	385	gi|9626180|	805	331	1.00E-89	hexon assembly-associated protein
	25472..25774	100	gi|9626578|	233	129	9.00E-30	pVIII protein
	27021..27494	157	gi|1279435|	166	314	7.00E-85	HLA-binding protein
	29892..30287	131	gi|6940696|	130	264	4.00E-70	E3B protein
	30280..30672	130	gi|6940697|	130	272	1.00E-72	E3B protein
	complement(30770..30919)	49	gi|9626584|	53	54.3	2.00E-07	U protein
	complement(32308..32970)	220	gi|3913555|	292	464	e-130	E4 protein
	complement(33116..33478)	120	gi|1699394|	120	259	2.00E-68	E4 protein
	complement(33481..33834)	117	gi|1699393|	117	243	7.00E-64	E4 protein
	complement(33831..34058)	75	gi|1699392|	130	142	5.00E-34	E4 protein
	complement(34266..34463)	65	gi|1699391|	125	132	7.00E-31	E4 protein

Table 2.5 (continued)

Group	Prediction	Predicted Length	Best BLASTP hit	BLASTP length	Score	E-value	Annotated Function
	Human herpesvirus 3, complete genome.		**NC_001348**				
	10678..10905	75	gi|13242466|	87	112	9.00E-25	membrane protein
	Human herpesvirus 4, complete genome.		**NC_001345**				
	503..805	100	gi|330387|	365	160	5.00E-39	latent membrane protein
	1546..1680	44	gi|330387|	365	85.8	7.00E-17	latent membrane protein
	166576..166920	114	gi|126379|	497	257	4.00E-68	latent membrane protein
	complement(169031..169474)	147	gi|126373|	386	224	6.00E-58	latent membrane protein
	Human herpesvirus 5, complete genome.		**NC_001347**				
	160003..160173	56	gi|7542409|	176	97.1	3.00E-20	interleukin-10-like protein
	Human herpesvirus 6B, complete genome.		**NC_000898**				
	23343..23774	143	gi|11346494|	305	300	1.00E-80	G-protein coupled receptor
	Human herpesvirus 7, complete genome.		**NC_001716**				
	129708..129848	46	gi|2746315|	153	101	2.00E-21	membrane glycoprotein
	Human papillomavirus type 1a, complete genome.		**NC_001356**				
	812..2650	612	gi|137646|	612	1251	0	replication protein E1
	Human papillomavirus type 53, complete genome.		**NC_001593**				
	892..1140	82	gi|9627323|	631	125	1.00E-28	Replication protein E1
	1391..1591	66	gi|9627323|	631	104	2.00E-22	Replication protein E1
	Human papillomavirus type 56, complete genome.		**NC_001594**				
	895..1149	84	gi|9628585|	630	112	1.00E-24	Replication protein E1
	1395..2804	469	gi|9628585|	630	927	0	Replication protein E1
	Human papillomavirus type 71, complete genome.		**NC_002644**				
	559..828	89	gi|1491685|	100	99.8	8.00E-21	transforming protein E7
	3004..3858	284	gi|9626037|	383	264	1.00E-69	regulatory protein E2
	4443..5783	446	gi|13186281|	524	583	e-165	minor capsid protein L2
	5776..7341	521	gi|3845719|	505	689	0	late major capsid protein L1
	Macaca mulatta rhadinovirus, complete genome		**NC_003401**				
	70403..70888	161	gi|13506781|	234	279	2.00E-74	bZIP transcription factor
	71468..72160	230	gi|13506783|	275	292	4.00E-78	glycoprotein R8.1
	Murine adenovirus type 1, complete genome.		**NC_000942**				
	2897..3175	92	gi|209749|	97	187	2.00E-47	early E1A protein
	complement(29726..30076)	116	gi|9800520|	810	67.9	6.00E-11	tropoelastin
	Ovine papillomavirus 1, complete genome.		**NC_001789**				
	747..2624	625	gi|9627078|	611	744	0	replication protein E1
	2611..3780	389	gi|9627069|	416	379	e-104	regulatory protein E2
	3780..3941	53	gi|137747|	44	66.3	5.00E-11	transforming protein E5
	4268..5623	451	gi|9627086|	447	445	########	minor capsid protein L2
	Ovine papillomavirus 2, complete genome.		**NC_001790**				
	745..2628	627	gi|9627078|	611	753	0	replication protein E1
	2615..3778	387	gi|9627069|	416	369	e-101	regulatory protein E2
	3778..3930	50	gi|137747|	44	65.2	1.00E-10	E5 protein
	4122..5615	497	gi|9627086|	477	525	e-148	minor capsid protein L2
	Tupaia herpesvirus, complete genome.		**NC_002794**				
	complement(60731..61684)	317	gi|9845327|	478	120	2.00E-26	US22 family protein
	Vaccinia virus, complete genome.		**NC_001559**				
	complement(5422..5526)	34	gi|3096964|	351	67.1	3.00E-11	TNF receptor II
	complement(6231..6377)	48	gi|3096965|	586	96.7	4.00E-20	K1R protein (ankyrin repeat protein)
	76530..76721	63	gi|11346541|	63	130	3.00E-30	RNA polymerase
	162151..162264	37	gi|401315|	193	58.9	9.00E-09	guanylate kinase
	183524..183640	38	gi|3096966|	672	66.7	4.00E-11	D4L protein (ankyrin repeat protein)
	185397..185507	36	gi|3096965|	586	70.6	3.00E-12	K1R protein (ankyrin repeat protein)
	186212..186316	34	gi|3096964|	351	67.1	3.00E-11	TNF receptor II
ssDNA	**Chloris striate mosaic virus, complete genome.**		**NC_001466**				
	complement(1864..2376)	170	gi|137410|	295	348	3.00E-95	replication-associated protein
	Periplaneta fuliginosa densovirus, complete genome.		**NC_000936**				
	complement(5134..5388)	84	gi|5689346|	291	83.1	6.00E-16	structural protein
phage	**Bacteriophage bIL311, complete genome.**		**NC_002670**				
	2252..2464	70	gi|15673928|	68	139	4.00E-33	ps3 protein 14-like transcriptional regulator
	Bacteriophage L5, complete genome.		**NC_003695**				
	2..340	112	gi|4098413|	348	216	8.00E-56	integrase

Table 2.5 (continued)

Group	Prediction	Predicted Length	Best BLASTP hit	BLASTP length	Score	E-value	Annotated Function
	Bacteriophage lambda, complete genome.		**NC_001416**				
	34482..35036	184	gi|140702|	183	374	e-103	superinfection exclusion protein B
	complement(46459..46752)	97	gi|137520|	97	196	5.00E-50	Bor protein precursor
	complement(47042..47575)	177	gi|16128541|	150	309	2.00E-83	putative envelope protein
	Bacteriophage VT2-Sa provirus, complete genome.		**NC_000902**				
	complement(11467..11595)	42	gi|15830439|	217	59.7	5.00E-09	c1 repressor protein
	Chlamydia phage phiCPAR39, complete genome.		**NC_002180**				
	1..147	48	gi|9634956|	84	104	2.00E-22	nonstructural protein
	4425..4532	35	gi|9634956|	84	75.3	1.00E-13	nonstructural protein
	Enterobacteria phage HK022 virion, complete genome.		**NC_002166**				
	19015..20130	371	gi|9634179|	321	270	2.00E-71	tail fiber protein
	complement(26155..26307)	50	gi|9634191|	50	104	1.00E-22	kil protein
	32436..33047	203	gi|15832758|	188	106	3.00E-22	endonuclease
	33876..34316	146	gi|9910800|	146	294	4.00E-79	Protein Nin B
	35667..36029	120	gi|9634210|	120	244	4.00E-64	holiday-junction resolvase
	Enterobacteria phage Mu, complete genome.		**NC_000929**				
	complement(33531..34064)	177	gi|96899|	177	360	8.00E-99	tail fiber assembly protein
	complement(34067..35053)	328	gi|96901|	536	678	0	tail fiber
	Roseophage SIO1, complete genome.		**NC_002519**				
	complement(39527..39826)	99	gi|9964612|	271	124	3.00E-28	gp5-like protein
	Streptococcus thermophilus bacteriophage 7201, complete g		**NC_002185**				
	3148..3330	60	gi|9634634|	218	116	3.00E-26	Erf protein
	Streptococcus thermophilus bacteriophage Sfi21, complete g		**NC_000872**				
	37175..37687	170	gi|9635004|	167	317	6.00E-86	DNA binding protein
	Sulfolobus Virus 1, complete genome.		**NC_001338**				
	12585..13001	138	gi|75696|	144	270	7.00E-72	structural protein VP1
Retroid	**Abelson murine leukemia virus, complete genome.**		**NC_001499**				
	4425..4580	51	gi|332031|	636	104	2.00E-22	env polyprotein
	Feline immunodeficiency virus, complete genome.		**NC_001482**				
	9006..9170	54	gi|128015|	122	118	1.00E-26	nef protein
	Friend spleen focus-forming virus, complete genome.		**NC_001500**				
	2173..2292	39	gi|11120675|	1733	79.6	5.00E-15	gag polyprotein
	2289..2543	84	gi|510896|	538	168	2.00E-41	gag polyprotein
	Human foamy virus, complete genome.		**NC_001736**				
	11054..11827	257	gi|227764|	356	562	e-159	bel-2 protein
	Human T-cell lymphotropic virus type 2, complete genome.		**NC_001488**				
	6..119	37	gi|6539751|	48	77.6	2.00E-14	tax protein
	Moloney murine sarcoma virus, complete genome.		**NC_001502**				
	2485..2967	160	gi|9626961|	1737	271	5.00E-72	pol polyprotein
	2945..3388	147	gi|9626961|	1737	293	1.00E-78	pol polyprotein
	4563..4718	51	gi|332031|	636	102	5.00E-22	envelope protein
	Murine osteosarcoma virus, complete genome.		**NC_001506**				
	complement(2305..2706)	133	gi|15822914|	137	250	4.00E-66	ubiquitin-like protein
	Murine sarcoma virus, complete genome.		**NC_001363**				
	2970..3452	160	gi|9626961|	1737	271	5.00E-72	pol polyprotein
	3430..3873	147	gi|9626961|	1737	293	1.00E-78	pol polyprotein
	5048..5203	51	gi|332031|	636	102	5.00E-22	spike protein
	Simian foamy virus, complete genome.		**NC_001364**				
	3..377	124	gi|9626108|	417	279	1.00E-74	bet protein
	Simian immunodeficiency virus, complete genome.		**NC_001549**				
	3..335	110	gi|9627209|	223	247	5.00E-65	nef protein
	Simian type D virus 1, complete genome.		**NC_001551**				
	5194..5973	259	gi|9627214|	1771	450	e-125	pol polyprotein
	Y73 sarcoma virus, complete genome.		**NC_001404**				
	2865..3194	109	gi|13508442|	611	206	7.00E-53	transmembrane envelope protein
	Barmah Forest virus, complete genome.		**NC_001786**				
	5679..7298	539	gi|7444406|	2493	816	0	nonstructural polyprotein
ssRNA(+)	**Northern cereal mosaic virus, complete genome.**		**NC_002251**				
	6740..12916	2058	gi|2961429|	1967	536	e-150	polymerase

Findings in Individual Genomes

A closer look on several new gene predictions is now taken. As mentioned above an *ab initio* approach can identify genes missed by a similarity search. Still, these genes and encoded proteins could successfully be functionally characterized since the set of objects to work with is rather narrow and it is possible to concentrate all available resources for in depth analysis of these putative proteins.

In the well studied genome of bacteriophage λ (Sanger *et al.*, 1982, Accession JO2459), as many as 5 new genes were identified. By now these genes have already been included in the latest RefSeq version of the phage λ annotation (NC_001416). Two genes, coding for a putative envelope protein (NP_597781) and Bor protein precursor (NP_597780), turned out to be similar to genes in prophage CP-933X, being a part of the *E. coli* O157 genome (NC_002655). A gene for superinfection exclusion protein B (NP_597779) must have been known for some time since its protein product had been included into the PIR database (P03762). At present, the other two genes are classified as hypothetical.

In another well studied virus, *Variola virus* (also known as SmallPox), over 20 new genes were identified. Many of these have similarity with other poxviruses, such as a gene detected at 6694..6870 on the complementery strand which has identical regions with genes from CamelPox, MonkeyPox, and CowPox. Another discovered gene, 7038..7313 on the complementary strand, also has high similarity to genes not only in other pox viruses but also in other strains of *Variola*. The RefSeq record for this virus was last updated in January 2003, and it is curious as to why the genes with protein

products highly similar to proteins in other poxviruses have not been listed in the public databases as of yet.

In *Porcine adenovirus A* (NC_001997), 16 new genes were identified by the *ab initio* predictions corroborated by similarity search. For instance, the protein encoded by predicted ORF # 6 is a member of a family of DNA polymerases present in 39 other adenoviruses.

A potentially important finding made with help of GeneMarkS is a gene located in positions 10443-11138 of the genome of *Alcelaphine* herpesvirus *1* (NC_002531) and carrying a code for a 232aa-long putative protein (NP_597933). Initially, the new protein was shown to be similar to the uncharacterized putative protein ORF E4 (NP_042601, AAC13792) of unclassified γ-herpesvirus *Equine* herpesvirus *2*. A subsequent PSI-BLAST search revealed statistically significant similarity between these two proteins and recently discovered antagonists of the lymphocryptovirus antiapoptotic BCL-2 proteins (Bellows *et al.*, 2001). Later, the sequence of the third non-lymphocryptovirus protein, hypothetical v-BCL2 of another unclassified γ-herpesvirus (porcine lymphotropic herpesvirus 1) was released (Goltz *et al.*, 2002). It has been shown that its sequence is very similar to the newly predicted protein. The PSI-BLAST search profile built from the three proteins further identified similarity (P-value = 0.0008), with ORF1 protein of *Callitrichine* herpesvirus *3* (a lymphocryptovirus BALF1-like BCL-2 like protein) and with the BALF1 protein (AAK01916) of *Allitrichine* herpesvirus 3 (a lymphocryptovirus) with the P-value 0.007. The output of the third iteration of PSI-BLAST included all the BALF1-like proteins at the top of the list. Human GRS protein and other BCL-2 like non-viral proteins were also present in the list at a substantial score distance.

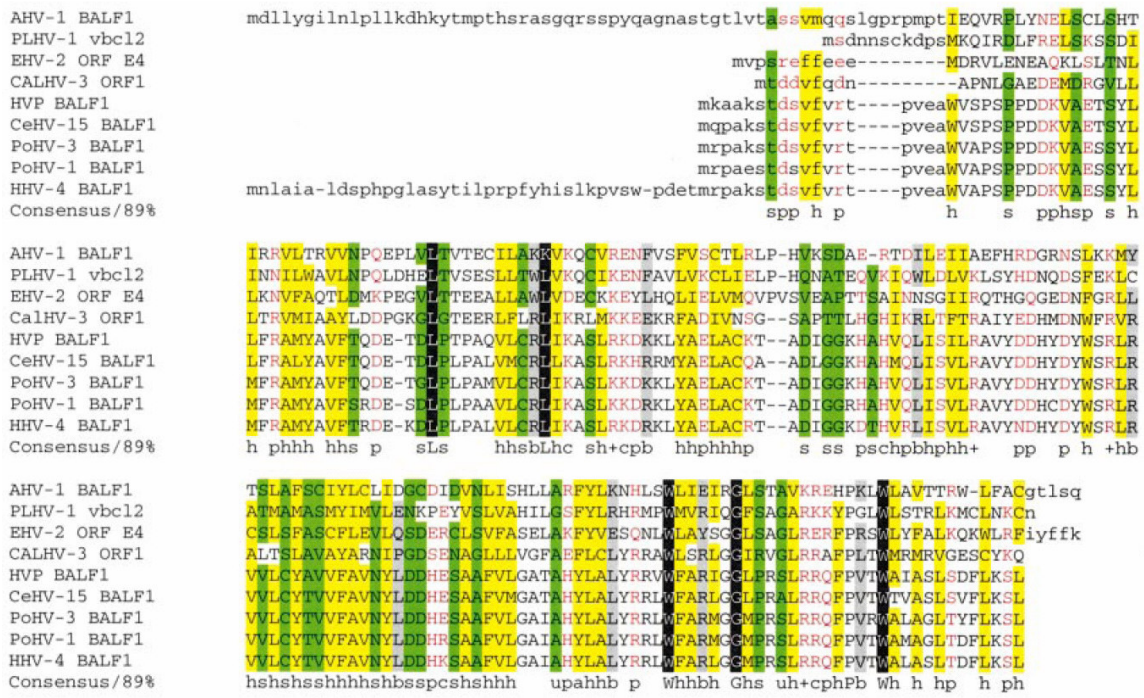

Figure 2.5 - MultAlin alignment of (putative) BALF1-like proteins. The variable N- and C-termini are shown in lower case. Protein names are abbreviated as follows: AHV-1 BALF1, BALF1 homolog (NP_597933) predicted by GeneMarkS in the genome of Alcelaphine herpesvirus 1 (NC_002531); PLHV-1 vbcl2, Porcine lymphotropic herpesvirus 1 hypothetical v-bcl2 (AAM22111); CALHV-3 ORF1, Callitrichine herpesvirus 3 ORF1 (AAK38208); HVP BALF1, Herpesvirus papio BALF1 (AAK01916); PoHV-3 BALF1, Pongine herpesvirus 3 BALF1 (AAK60342); HHV-4 BALF1, Human herpesvirus 4 BALF1 (NP_039912); PoHV-1 BALF1, Pongine herpesvirus 1 (AAK01917); CeHV-15 BALF1, Cercopithicine herpesvirus 15 (AAK95480); EHV-2 ORF E4, Equine herpesvirus 2 ORF E4 protein (NP_042601). The conserved positions are color coded based on the type of amino acid residue as indicated in the consensus line, where h and a stand for hydrophobic residues (A, C, F, I, L, M, V, W, Y: yellow background in alignment) and for aromatic residues (F, Y, W), respectively; b stands for `large ' residues (E, K, R, I, L, M, F, Y, W: gray background); p stands for polar residues (D, E, H, K, N, Q, R, S, T: shown in pink); s and u stand for small residues (A, C, S, T, D, N, V, G, P: green background) and tiny residues (G, A, S), respectively; c and + stand for charged residues (K, R, D, E, H: shown in pink) and positively charged residues (K, R), respectively. Invariant amino acid residues (in 85% or more sequences) are highlighted with black background.

The RPS-BLAST search immediately detected a BCL-motif in all three non-lymphocryptovirus proteins. Figure 2.5 shows an alignment (Corpet, 1988) of the newly predicted protein (NP_597933) with proteins NP_042601, AAM22111 and all the lymphocryptovirus BALF1 proteins. Interestingly enough, the TBLASTN search failed to reveal additional un-annotated homologs of NP_597933. It is tempting to speculate that, given the function of BALF1 (Bellows *et al.*, 2002), the newly identified BALF1-like protein may be involved in a complex regulation of the host cell apoptosis, presumably as antagonist of the herpesvirus antiapoptotic BCL-2 proteins, and, perhaps, is a part of a gene network involved in cancerogenesis.

Another interesting new finding was a gene predicted by GeneMarkS (ORF #65) in the genome of the Epstein-Barr virus (HHV-4, NC_001345). Initially, the protein product of this gene was found to be significantly similar (with P-value < 10^{-5}) to uncharacterized ORF26/ORF35 proteins of other γ-herpesviridae. The subsequent PSI-BLAST search has revealed after four iterations a similarity (with P-value = 0.0006) to the ORF26/ORF35 protein family and the ORF48 protein of Equine herpesvirus 4, an α-herpesvirus. The ORF48 protein belongs to the UL14 family of proteins which are present in a minor component of the virion tegument and possess heat shock protein-like functions (Yamauchi *et al.*, 2002). Eight further PSI-BLAST iterations brought up all the members of this family. Multiple alignment of the ORF26/ ORF35 and UL14-like protein sequences (Figure 2.6) highlights common features that could not be readily seen in pair-wise alignments, particularly, similar patterns of distribution of charged residues (coloring as described above). The observed sequence similarity strongly indicates a common function which remains to be determined by direct experiments. It is likely that

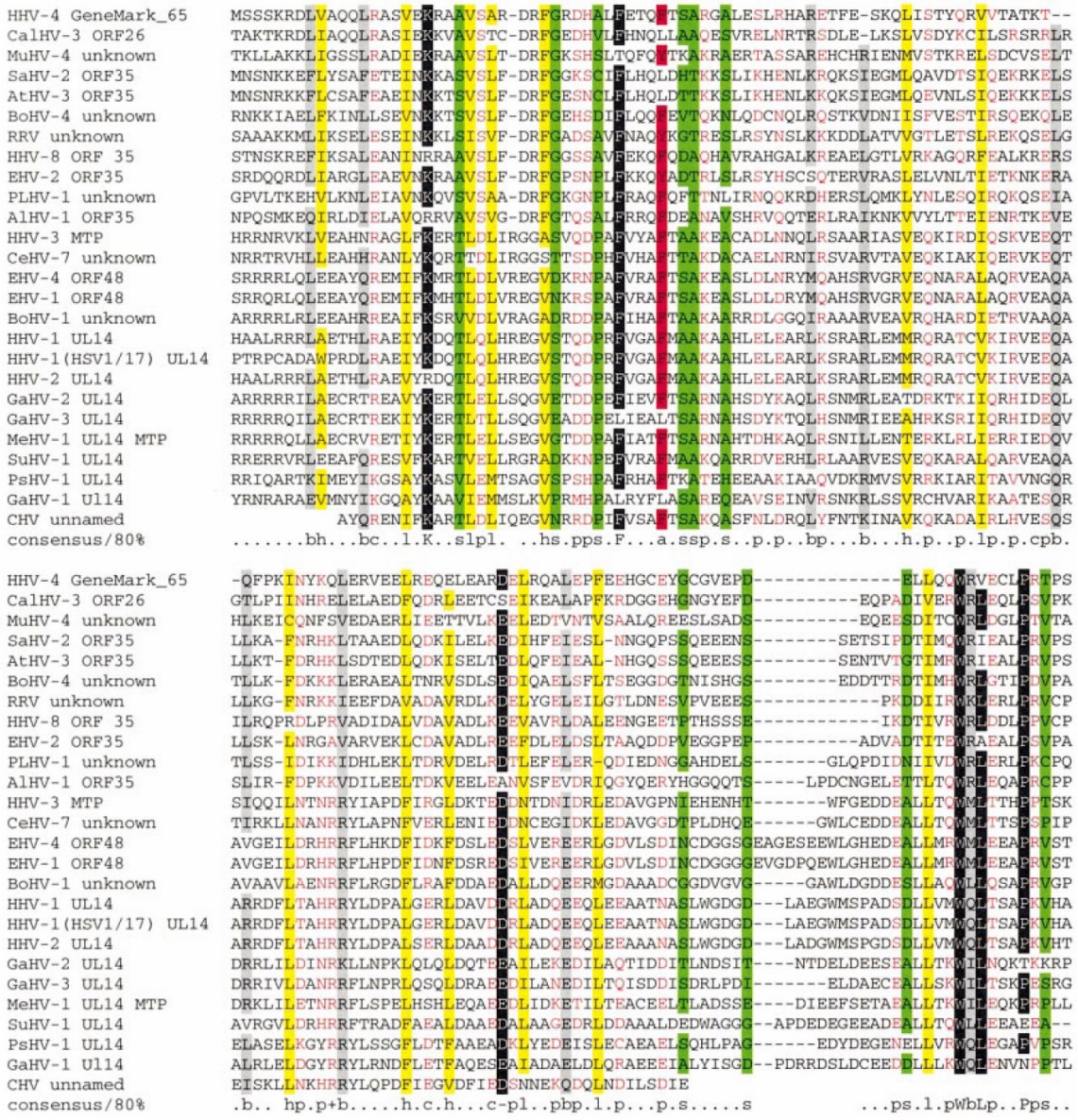

Figure 2.6 – Alignment of the sequences of ORF26/ORF35 and UL14-like proteins. For most sequences, the N- and C-termini are not shown. The coloring is as in Figure 2.4.

these proteins play an important role, since the members of the ORF26/ORF35 protein family are now confirmed to be present in all complete genomes of γ-herpesviruses. Interestingly, none of the β-herpesviruses genomes has a TBLASTN detectable homolog of ORF26/ORF35 or UL14, which indicates that ORF26/ORF35 proteins are likely to fulfill a subfamily-level function.

Even annotation of recently sequenced viral genomes can be improved through statistical means. The analysis of the SARS coronavirus annotation (April 14, 2003) showed a gene on the direct strand at 27849..28103 with a high protein-coding potential that was missed in the annotation. This gene was included with a later annotation (May 15, 2003), although based on other type of evidence. Still, had researchers interested in this genome merely viewed the results of the statistical analysis of the virus done by our method, they might have elucidated this gene much sooner.

Some coding regions in viral genomes were missed in the earlier annotation because of their unusual organization. For instance, some viral genes contain a weak, read-through stop codon, which in the original annotation is considered to be the end of the gene; thus, a part of the real gene (and protein) is missed. For example, a GeneMarkS prediction (ORF #2) in *Barmah Forest virus,* recovers the second part of the non-structural polyprotein gene in positions 5679-7298, the one missed in the original record U73745. Only after combining together these two parts, the resolved protein (NC_001786) shows full length similarity to the complete polyprotein encoded, for instance, in *Ross River virus*.

The vast majority of genes in viral genomes have no introns. Still, there are a few genes with introns and even with introns engulfing whole separate genes located inside

introns, such as an IE glycoprotein gene, HCMVUL37, in human herpesvirus *5* (NC_001347). Genes interrupted by introns are identified by GeneMarkS as series of separate protein-coding genes. For instance, in Enterobacteria phage T4 a gene for DNA topoisomerase small subunit protein (NC_000866) consists of two exons both predicted by GeneMarkS as separate genes. Developing an *ab initio* approach for exact prediction of viral introns is a challenging problem. However, quite frequently the combination of the intrinsic and extrinsic evidence provides sufficient information for further delineation of exon-intron structure by manual intervention. For instance, in the complete genome of Human adenovirus D (Human adenovirus type 17), GeneMarkS revealed 32 potential genes or gene fragments missed in the original annotation (AF108105). Only 11 of them appeared to be complete genes while the other 21 predicted coding regions were manually assembled into 9 genes in the RefSeq record (NC_002067).

For many predicted proteins no significant similarity was found to any known protein in the NCBI protein database. However, the absence of such similarity does not rule out the possibility that these predictions are real genes.

Discussion

Gene finding in small genomes such as viruses can be a daunting task for even the most sophisticated statistical methods. Still, by improving existing algorithms and model building techniques, a higher degree of accuracy can be achieved than by using arbitrary rules such as the annotation of every ORF longer than some predetermined size as a gene. This is especially important in the finding of small genes, which can be easily overlooked or ignored.

A statistical method for analyzing prokaryotic and eukaryotic viruses has been presented in this chapter. This procedure was applied to over 5000 complete viral genome records in GenBank, and the resulting annotations were compiled into an interactive database, VIOLIN. This database allows the user to peruse genes predicted by the statistical analysis and to use the protein-specific links to attempt to further characterize those genes. An in-depth study was conducted on a subset of newly predicted genes in the database, with many new and interesting facts being elucidated.

As more genomic sequences for viruses and phages are determined, the ability to cross reference proteins of one virus with one or more viruses of similar type would become very useful. To this end, the future direction of the VIOLIN database development is the implementation of intra-database comparative genomic tools that would allow for searching for genes, proteins, and regulatory regions that are shared between several genomes. The VIOLIN database can also be updated to reflect the improvements made in the viral gene prediction algorithm, including the development of better methods of predicting the exon/introns structure.

CHAPTER 3

COMPUTATIONAL GENOMIC ANALYSIS OF THE GENOME OF THE HERPES B VIRUS (CERCOPITHECINE HERPESVIRUS 1) FROM A RHESUS MONKEY

Introduction

The complete DNA sequence of the genome of herpes B virus (Cercopithecine herpesvirus 1) strain E2490, isolated from a rhesus macaque, was determined. The total genome length is 156,789nt, with 74.5% G+C composition. The overall genome organization is characteristic of alphaherpesviruses. Unexpectedly, B virus lacked a homolog of the HSV gamma(1)34.5 gene, which encodes a neurovirulence factor. Absence of this gene was verified in two low-passage clinical isolates derived from a rhesus macaque and a zoonotically infected human, as well as through statistical profile searching. This finding suggests that to sustain efficient replication in neuronal cells B virus most likely utilizes mechanisms distinct from those of HSV. Despite the considerable differences in G+C content of the macaque and B virus genes (51% and 74.2%, respectively), codons used by B virus are optimal for the tRNA population of macaque cells. In collaboration with the sequencing group, we built specialized statistical models for this viral genome and improved the current annotation. We also derived models, HMM profiles and position specific weight matrices for HSV1 and HSV2 gamma(1)34.5 genes and used them to search the genomic sequence for the missing gene. As these attempts did not bring in a positive result, several hypotheses were proposed on how this genome functions without a gamma(1)34.5 homolog

This project (Perelygina *et al.*, 2003) was done in collaboration with the Hilliard group at Georgia State University. The bench work and sequencing presented in this

section was conducted by the Georgia State group, while the gene characterization, comparative genomics and statistical analyses were done at Georgia Tech.

Methodology

The B virus laboratory strain E2490 and two B virus clinical isolates were obtained from the National B Virus Resource Laboratory (Atlanta, Georgia) and sequenced at Georgia State University. Identification of open reading frames (ORFs), repeats, and DNA regulatory sequences was performed with DNAStar suite of programs. GenBank database searches were carried out by BLASTN, BLASTP, and BLASTX with default settings. Multiple alignments between B virus, HSV-1, and HSV-2 genes and proteins were performed by DNAStar MegAlign program (version 4.0.3) using the PAM250 amino acid substitution matrix and Joltun Hein method (Hein, 1990) with the following parameters: gap opening penalty = 11 and gap extension penalty = 3. Pairwise sequence identity values were determined from the generated alignments.

The GeneMark and GeneMark.hmm gene-finding programs were used to refine the gene annotation (Borodovsky and McIninch, 1993; Lukashin and Borodovsky, 1998). The parameters of the statistical models were defined by training on the set of B virus ORFs (majority of the B virus genome ORFs) encoding protein products homologous to known HSV-1 and HSV-2 proteins. These statistical methods are able to identify frameshifts and unusual gene starts in addition to detecting genes additional to those found by similarity search. To prove that a homolog of the HSV neurovirulence factor is absent in the B virus genome, BLASTP and PSI-BLAST searches with gamma(1)34.5 protein as the query were used against the whole B virus genome translated in six frames.

In addition, a HMM profile was created for the conservative domain of the ICP34.5 protein and used to scan the translated B virus genome.

In the search for unique genes, the PSI-BLAST and RPS-BLAST programs were used to search for conserved domains in proteins predicted by statistical methods. Predicted proteins shorter than 20 residues were excluded from the analysis. Analyses of protein structures were performed using the DNAStar Protean program (version 4.0.3), signal peptide prediction program SignalP version 2.0.b2 (http://www.cbs.dtu.dk/services/SignalP-2.0/), and transmembrane prediction programs TMHMM version 2.0 (http://www.cbs.dtu.dk/services/TMHMM/), DAS (http://www.sbc.su.se/~miklos/DAS/), and TMpred (http://www.ch.embnet.org/software/TMPRED_form.html).

The annotation of the B virus genome (GenBank accession number AF533768) along with its completed sequence has been deposited to the NCBI database.

Results

The B virus genome sequence overview

The complete B virus genomic sequence was assembled according to the HSV-1 prototype genome structure (Roizman and Knipe, 2001). The genome of B virus is 156,789nt in length, which is considerably shorter than it was predicted earlier based on the constructed B virus physical map (Harrington *e al.*, 1992) and direct measurements by electron microscopy (Ludwig *et al.*, 1983). The 74.5% G+C base composition of the genome is very similar to earlier estimates of 75% determined by DNA buoyant density centrifugation (Ludwig *et al.*, 1983). Table 3.1 presents a comparison of the length and G+C content of B virus, HSV-1, and HSV-2 by genomic regions. The G+C composition

was observed to be elevated in all B virus regions compared to that in HSV types 1 and 2, with unique B virus regions demonstrating the highest G+C content among known herpesviruses. The sizes of genomic segments are comparable among the viruses with the exception of the RS segment, which in the B virus genome is longer by 1.5 kb.

Although overall sequence homology between the sequences of B virus, HSV-1, and HSV-2 is low, the sequence composition of the cleavage and packaging signals was very well conserved. An A-rich motif followed by three CGCGGCG motifs formed the B virus pac2 signal. Interestingly, most herpesviruses have only one copy of the CGCGGCG motif, which contributes significantly to the efficiency of genome cleavage and packaging (McVoy *et al.*, 1998). The B virus pac1 contains three conserved motifs, an A-rich region flanked by two stretches of seven C residues and 13 C+G residues. Both pac1 and pac2 are located at conserved distances from the genome ends (Deiss *et al.*, 1986), 31nt from the L terminus and 34nt from the S terminus of B virus DNA. There are no DR2 repeats between the pac1 and pac2 signals in the B virus a sequence, which are thus noticeably shorter than such a sequence HSV-1.

Ten sets of tandem duplications were identified in the B virus genome. Genomic locations of sets, numbers of repeat units in each set, sizes, and sequences are given in Table 3.2. Most duplications were present at locations compatible with the locations of HSV-1 duplications. Six sets were present in two copies in the genome due to location in the terminal and internal copies of the RS and RL segments. Only two repeat sets were located in protein-coding regions, the UL36 and US4 (gG) genes.

Both origins of the B virus DNA replication, oriL and oriS, were tandemly duplicated and present in locations corresponding to HSV oriL and oriS locations.

Table 3.1 - Comparison of B virus, HSV-1, and HSV-2 genomic regions

Region	Virus	Length (bp)	% G+C
U_L	B virus	107,815	72.9
	HSV-1	107,947	66.9
	HSV-2	109,689	68.9
U_S	B virus	14,687	73.2
	HSV-1	12,980	64.3
	HSV-2	14,329	66.2
TR_L	B virus	9,021	79.4
	HSV-1	9,912	71.6
	HSV-2	9,297	74.4
TR_S	B virus	8,234	80.4
	HSV-1	6,677	79.5
	HSV-2	6,711	80.1
Whole genome	B virus	156,789	74.5
	HSV-1	152,261	68.3
	HSV-2	154,746	70.4

Table 3.2 - Sets of repeated sequences in the B virus genome

Set	Genomic region	Location of set		Sequence of repeat unit	Unit size (bp)	No. of units	Comments
		First copy	Second copy[a]				
1	R_L	824–1022	124836–125034	GGGGGTCCTGGGGGTCCGGGGTCGCC	26	8	
2	R_L	2244–2315	123543–123614	GGGGGTCTC	9	8	Located in ICP0 intron
3	R_L	3658–3729	122129–122200	GCCCGGCGCCCAAGTCCC	18	4	
4	R_L	5164–5244	120614–120694	CCAGAAGCAGAGAGGGGCGGGGGCTCC	27	3	
5	U_L	43039–43123		GGGGGTGCGGGGGCGGT	17	5	Located between UL21 and UL22
6	U_L	71559–71756		GCAGGGGCA	9	22	Encodes PAA repeat in UL36
7	U_L	115984–116238		CCCCCCTCCCCTCCCCCGCG	20	13	
8	R_S	131851–132461	149963–150573	CCCTTCCCCCCTT	13	47	Flanks *oriS*
9	R_S	133448–133785	148639–148976	CCCCCGCGCACCCCTCGCCCTCCCCT	26	13	Flanks *oriS*
10	U_S	139337–139444		CCCGCCCCCACCACCACC	18	6	Encodes PAPTTT repeat in gG

[a] A location of the second copy of a set is given for reiteration sets from the R_S and R_L genomic regions.

Thus, six origins of DNA replication exist: two copies of oriL (oriL1 and oriL2) in the UL region and two copies each of oriS (oriS1 and oriS2) in the terminal and internal RS regions. Duplications of either the oriL or oriS sequence were also found in HSV-1 strain ANG and HSV-2 strain HG52, respectively (Dolan *et al.*, 1998; Gray and Kaerner, 1984), but concurrent duplication of both origins has not been described previously for herpes simplex viruses. All B virus origins share the core element of a 94-bp perfect palindrome (Figure 3.1) containing two predicted binding sites for the origin-binding protein (OBP), box I, and box III (Elias *et al.*, 1990; Hazuda *et al.*, 1991; Olivo *et al.*, 1988). Furthermore, the nucleotide sequences of the B virus oriS and oriL core elements were extremely conserved and almost identical to the HSV oriL but differed from HSV oriS (Hardwicke *et al.*, 1992; Martin *et al.*, 1991). Whether the existence of six nearly identical origins of replication in the B virus genome has any functional significance or is just an artifact due to repeated passage of the virus in cell culture currently remains unknown.

Gene and protein identification

Both major approaches to gene identification, extrinsic and intrinsic, were used to identify and characterize B virus genes. Extrinsic methods, such as BLASTP identify protein-coding genes by detecting similarity of translated protein sequences to the primary structure of a known protein. The intrinsic approach, an *ab initio* statistical method such as GeneMark, identifies protein-coding regions by detecting specific frequency patterns in nucleotide order, including the codon usage pattern. These two types of methods have complementary strengths in terms of sensitivity and specificity.

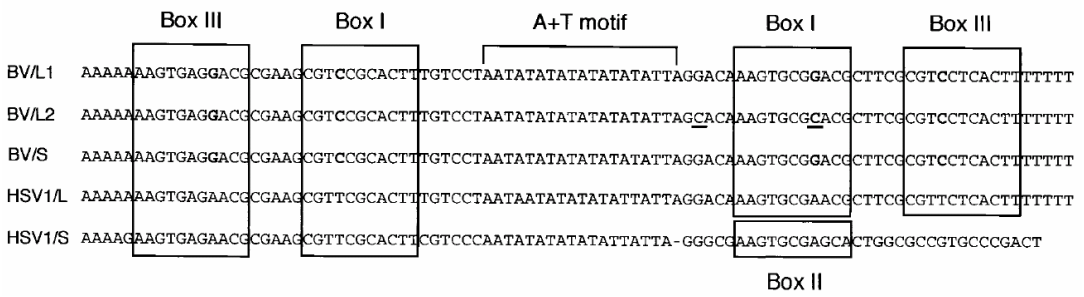

Figure 3.1 - Comparison of B virus and HSV-1 origins of replication. The DNA sequences of the *oriL* and *oriS* core elements are shown. BV/L1, B virus oriL1; BV/L2, B virus oriL2; BV/S, B virus *oriS1* and *oriS2*; HSV1/L, HSV-1 oriL; HSV1/S, HSV-1 oriS. The OBP-binding sites (box I, boxII, and box III) are boxed. Residue substitutions in all B virus origins relative to HSV-1 *oriL* are shown in bold. Residue substitutions in B virus oriL2 relative to *oriL1* are underlined.

The B virus genomic sequence can be efficiently analyzed by similarity search methods due to extensive knowledge of the closely related viral species HSV-1 and HSV-2. Still, unique virus-specific genes could be missed. Therefore, in addition to the similarity search, we used the *ab initio* methods, GeneMark and GeneMark.hmm, trained on the set of genes confirmed by similarity search. These methods were modified to allow a noncanonical translation initiation codon, GTG (Medhi *et al.*, 1990), which could be used in this highly G+C-rich genome.

All but one B virus gene were identified on the basis of sequence homology to HSV-1 and HSV-2 genes and named correspondingly. Table 8.3 summarizes their locations in the genome, predicted sizes, and percent identities with the corresponding HSV-1 and HSV-2 gene products. Seventy-two genes existed as a single copy within unique genomic regions, whereas two genes, ICP0 and ICP4, appeared twice due to duplication of the large and small repeat regions where they reside. Two ORFs (RL2 and UL15) are predicted to have introns matching ones in the corresponding HSV genes.

It was inferred that the B virus UL1 and UL2 genes use the GTG codon as a start codon at genomic sequence positions 9072 and 9814, respectively. This prediction was made statistically and was corroborated by the presence of conserved regions upstream of the canonical ATG. Both predicted translational initiation GTG codons aligned with the ATGs of the HSV UL1 and UL2 genes. Conservation of the UL1 GTG start codon was confirmed by sequence analysis of the 5' region of the UL1 gene from the two B virus clinical isolates (data not shown).

Table 3.3 - ORFs and other features of the B virus genome

ORF or feature	Location		Strand	Length (codons)	Identity (%)		Characteristics and predicted function(s)[a] (reference)
	Start	Stop			HSV-1	HSV-2	
a sequence	1	223					Terminal direct repeat
TR_L	1	9021					Terminal copy of large repeat region
RL2			+	701	43.5	40.1	Immediate-early protein ICP0; multifunctional regulatory protein
Exon 1	2194	2241	+	16			
Exon 2	2461	3003	+	181			
Exon 3	3125	4636	+	504			
U_L	9022	116837					Unique large region
UL1	9072	9746	+	225	65.5	53.4	Virion membrane glycoprotein L; in complex with gH; membrane fusion
UL2	9814	10575	+	254	71.3	68.9	Uracil-DNA glycosylase; DNA repair
UL3	10711	11394	+	228	61.8	61.8	Colocolization with ICP22 and UL4 in small, dense nuclear bodies
UL4	12120	11509	−	204	56.1	56.5	Colocolization with ICP22 and UL3 in small, dense nuclear bodies
UL5	14820	12172	−	883	82.7	83.0	Component of helicase-primase complex
UL6	14819	16870	+	684	69.5	71.6	Capsid protein; DNA cleavage/packaging
UL7	16821	17711	+	297	65.2	65.2	Capsid protein; DNA cleavage/packaging
UL8	20171	17886	−	762	58.2	60.4	Component of helicase-primase complex
UL9	22847	20223	−	875	78.3	78.8	*ori* binding protein; helicase activity
UL10	22714	24132	+	473	65.2	65.1	Virion membrane glycoprotein M; proposed role in capsid envelopment
UL11	24733	24437	−	99	50.0	53.1	Myristylated tegument protein; capsid envelopment
UL12	26553	24679	−	625	67.8	69.1	DNase; endoexonuclease; processing of DNA replication intermediates
UL13	28097	26553	−	515	66.9	68.8	Virion protein kinase
UL14	28513	27869	−	215	65.6	66.5	Minor tegument protein
UL15			+	739	87.5	87.5	DNA cleavage/packaging; transiently associated with maturing capsids
Exon 1	28599	29624	+	342			
Exon 2	33199	34389	+	397			
UL16	30898	29795	−	368	64.2	66.7	Capsid associated; DNA cleavage/packaging; located in UL15 intron
UL17	33015	30919	−	699	66.8	69.5	Tegument protein; DNA cleavage/packaging
UL18	35574	34618	−	319	82.8	82.4	Capsid protein VP23; forms triplexes with VP19C that connect pentons and hexons in capsids
UL19	39918	35785	−	1,378	86.9	85.5	Major capsid protein VP5; forms pentons and hexons of capsid shell
UL20	40783	40112	−	224	64.6	66.4	Virion membrane protein; virion egress; *syn5* locus
UL21	41384	42964	+	527	66.2	68.3	Nucleotidylylated phosphoprotein; interacts with microtubules and facilitates intracellular transport of the virus (65)
UL22	45736	43193	−	848	59.1	60.9	Virion membrane glycoprotein H; in complex with gL; membrane fusion, entry, cell-to-cell spread
UL23	47107	45995	−	371	59.8	60.7	Thymidine kinase
UL24	47042	47842	+	267	66.4	66.4	Nonglycosylated membrane-associated protein; *syn5* locus
UL25	48036	49775	+	580	79.8	79.8	Minor capsid protein; DNA packaging; possible role in DNA anchoring (50)
UL26	49937	51763	+	609	65.7	65.4	Capsid maturation protease
UL26.5	50834	51763	+	310	58.8	57.0	Scaffolding protein
UL27	54815	52137	−	893	79.9	80.4	Virion membrane glycoprotein B; cell entry; contains *syn3* locus
UL28	57192	54835	−	786	84.0	84.8	DNA cleavage/packaging; transiently associated with maturing capsids
UL29	61257	57667	−	1,197	82.7	82.2	Single-strand DNA-binding protein; key role in assembly of DNA replication proteins
oriL1	61592	61789					Center of replication origin *oriL1*
oriL2	61795	61992					Center of replication origin *oriL2*
UL30	62173	65916	+	1,248	79.6	80.0	DNA polymerase catalytic subunit; complexes with UL42
UL31	66766	65861	−	302	80.9	81.3	Nuclear phosphoprotein; interacts with UL34; capsid egress from nucleus
UL32	68531	66759	−	591	74.2	74.5	DNA packaging; not associated with capsids
UL33	68530	68928	+	133	72.1	72.9	DNA packaging; not associated with capsids
UL34	68988	69803	+	272	66.8	71.3	Type II nuclear membrane-associated phosphoprotein; interacts with UL31; capsid egress from nucleus
UL35	69939	70283	+	115	51.8	47.3	Basic phosphorylated capsid protein VP26
UL36	80356	70490	−	3,289	61.2	62.2	Very large tegument protein; interacts with UL19 and UL37
UL37	84145	80624	−	1,174	69.4	68.7	Minor tegument protein
UL38	84598	85974	+	459	70.6	69.1	Capsid protein VP19C, forms triplexes with VP23 that connect pentons and hexons in capsids
UL39	86392	89385	+	998	66.7	66.0	Large subunit of ribonucleotide reductase
UL40	89434	90450	+	339	79.7	80.4	Small subunit of ribonucleotide reductase
UL41	92073	90619	−	485	72.3	73.0	Tegument phosphoprotein; virion-associated host shutoff (vhs) protein
UL42	92575	94002	+	476	48.6	48.8	Double-stranded DNA-binding protein, DNA polymerase subunit
UL43	94131	95270	+	380	44.4	49.1	Predicted membrane-associated protein
UL44	95527	96930	+	468	49.9	51.5	Virion membrane glycoprotein C; cell attachment; blocking host immune response
UL45	97166	97690	+	175	63.6	59.0	Type II membrane protein; possible role in cell fusion
UL46	100140	97978	−	721	57.7	58.9	Tegument phosphoprotein VP11/12; modulates alpha *trans*-inducing factor activity
UL47	102327	100267	−	687	59.9	58.5	Tegument phosphoprotein VP13/14; O-glycosylated; modulate α-TIF activity; RNA binding (60)
UL48	104246	102783	−	488	69.3	69.2	Major tegument protein VP16 (α-TIF); *trans*-activator of α genes
UL49	105449	104580	−	290	45.9	45.0	Major tegument protein VP22; binds RNA; carrier of mRNA from infected to uninfected cells (60)

Table 3.3 (continued)

ORF or feature	Location		Strand	Length (codons)	Identity (%)		Characteristics and predicted function(s)[a] (reference)
	Start	Stop			HSV-1	HSV-2	
UL49A	106092	105853	−	80	40.5	43.0	Envelope protein
UL50	106107	107216	+	370	55.9	54.3	Deoxyuridine triphosphatase
UL51	108064	107381	−	228	67.5	69.3	Capsid/tegument-associated phosphoprotein (15)
UL52	108126	111302	+	1,059	73.0	73.1	Component of helicase-primase complex
UL53	111254	112267	+	338	66.0	68.6	Membrane glycoprotein K; virion egress; contains *syn1* locus
UL53A			−	300			Hypothetical protein predicted by GeneMark and GeneMark.hmm
Exon 2	112186	112495	−	197			
Exon 1	112755	113344	−	103			
UL54	112644	114116	+	491	59.1	60.8	Immediate-early protein ICP27; regulates some early and all late gene expression
UL55	114429	115001	+	191	62.0	64.3	Nuclear matrix-associated protein
UL56	115834	115154	−	227	39.7	42.9	Type II membrane protein (38); involved in virus pathogenicity (35)
IR_L	116837	125634					Internal copy of large repeat region
RL2			−	701	43.5	40.1	Immediate-early protein ICP0; multifunctional regulatory protein
Exon 3	122733	121222	−	504			
Exon 2	123397	122855	−	181			
Exon 1	123664	123617	−	16			
a′ sequence	125635	125857					Inverted copy of *a* sequence
IR_S	125858	133868					Internal copy of small repeat region
RS1	131284	127724	−	1,187	65.0	66.8	Immediate-early protein ICP4; regulator of gene expression
Ori_s1	132795	132796					Center of replication origin *oriS1*
Ori_s2	132998	132999					Center of replication origin *oriS2*
U_s	133869	148554					Unique small region
US1	133900	135252	+	451	41.1	43.5	Immediate-early protein ICP22; required for optimal ICP0 expression
US2	136386	135478	−	303	56.1	54.0	Tegument protein
US3	136708	138084	+	459	60.8	61.1	Protein kinase; antiapoptotic activity
US4	138221	140242	+	674	29.2	39.2	Virion membrane glycoprotein G; entry into polarized cells
US5	140465	140830	+	122	34.0	26.6	Glycoprotein J; block apoptosis
US6	141296	142480	+	395	57.0	59.0	Virion membrane glycoprotein D; cell entry; interacts with cellular receptors
US7	142680	143885	+	402	46.1	51.0	Virion membrane glycoprotein I; in complex with gE; basolateral viral spread
US8	144253	145872	+	540	46.0	48.0	Virion membrane glycoprotein E; in complex with gI; basolateral viral spread
US8.5	145817	146185	+	123	42.3	45.9	Nucleolar phosphoprotein
US9	146309	146581	+	91	58.9	57.3	Tegument protein
US10	148139	147204	−	312	43.7	45.9	Tegument protein
US11	148290	147847	−	148	45.2	46.7	RNA-binding tegument protein; interacts with protein kinase R
US12	148555	148310	−	82	26.8	26.8	Immediate-early protein ICP47; inhibits antigen presentation
TR_s	148556	156566					Terminal copy of small repeat region
oriS2	149425	149426					Center of replication origin *oriS2*
oriS1	149628	149629					Center of replication origin *oriS1*
RS1	151140	154700	+	1187	65.0	66.8	Immediate-early protein ICP4; regulator of gene expression
a sequence	156567	156789					Terminal direct repeat

[a] Characteristics and functions of B virus proteins were deduced from the properties of known HSV-1 and HSV-2 homologs (31, 47, 57).

The extent of the amino acid identity between B virus and HSV polypeptides varied from 26.6% (US5) to 87.7% (UL15). The conservation of specific protein domains observed in HSV-1 and HSV-2 was mirrored when HSV-1 and B virus were compared. The following proteins were significantly conserved in B virus: DNA cleavage and packaging proteins, i.e., UL15, UL28, UL32, and UL33; capsid proteins, i.e., UL18, UL19, and UL38; proteins involved in DNA replication, i.e., UL2, UL5, UL29, and UL30; and glycoprotein B. The three least-conserved proteins in B virus were US4, US5, and US12. Similar levels of conservation have been described for homologous proteins in many other mammalian herpesviruses (Roizman and Pellett, 2001).

The B virus proteins could be divided into three groups based on the degree of similarity to HSV-1 and HSV-2 proteins. The largest group (46 proteins) showed greater similarity to HSV-2 proteins, e.g., DNA cleavage and packaging proteins. The second group (20 proteins), with capsid proteins among others, showed higher levels of similarity to HSV-1 proteins. In seven proteins, differences in similarity to HSV-1 or HSV-2 homologs were marginal, and these proteins formed the third group. The proteins that enable HSV-1 to replicate and reactivate more efficiently at orofacial sites and HSV-2 at genital sites are unknown, but the fact that B virus replicates and is reactivated with similar efficiencies at both sites may be explained by the similarities of selected B virus proteins with either HSV type 1 or 2 proteins. These hybrid properties in B virus raise new questions about alphaherpesvirus evolution.

In recent years, the existence of additional HSV-1-specific genes has been proposed and substantiated by experimental data. UL20.5 is located between UL20 and UL21, ORF P and ORF O map antisense to the 34.5 gene, and UL43.5 and UL27.5 map

antisense to UL43 and UL27, respectively (Chang *et al.*, 1998; Randall *et al.*, 1997; Randall and Roizman, 1997; Weigler, 1992; Whitley and Hilliard, 2001). Like HSV-2 (Dolan *et al.*, 1998), B virus has no equivalents to these proposed genes.

Determination of gamma(1)34.5 absence

One well-established gene, the HSV 134.5 (RL1) homolog, was observed to be absent in B virus after the initial gene identification screen. We conducted a more thorough search by creating a database of all open reading frames found in B virus and searching this database with the 134.5 homologs in both HSV 1 and 2. We were unable to find such a homolog in B virus.

The herpes simplex virus virulence factor gamma(1)34.5, the mouse myeloid differentiation protein MyD116, and the hamster growth arrest and DNA damage protein GADD34 share a 63-amino-acid carboxyldomain (Brown *et al.*, 1997), and it is thought that this particular domain is responsible for virulence. We constructed an HMM profile from these protein sequences (Figure 3.2) and searched the translated B virus genome sequence extensively. We were unable to find a high scoring sequence, suggesting that no protein posessing such a domain exists in B virus. Deletion of 134.5 leads to complete neuroattenuation of highly neurovirulent HSV-1 strains.

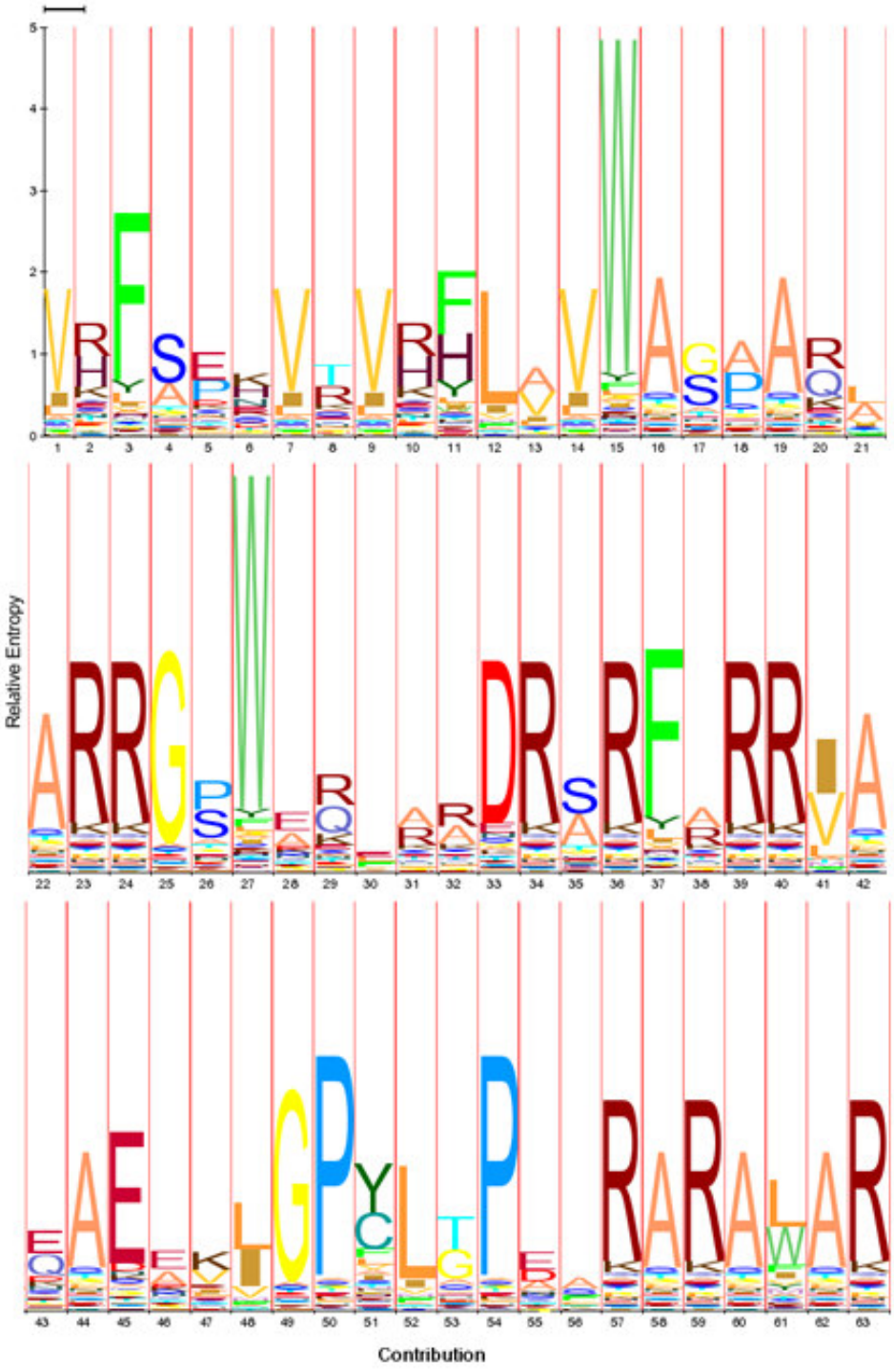

Figure 3.2– Sequence logo (Schuster-Boeckler *et al.*, 2004) of the HMM created from the ICP34.5, GADD34, and MyD116 neurovirulence domains in HSV-1, hamster, and mouse, respectively.

To determine that this observation was not limited to the laboratory strain of B virus, L-terminal SphI restriction fragments were cloned and sequenced from two low-passage clinical isolates, one of which was derived from a rhesus macaque (Perelygina *et al.*, 2003) and the other post mortem from a zoonotically infected human patient (Davenport *et al.*, 1994). Sequence comparison of these fragments with the corresponding fragment of the laboratory strain did not reveal any significant differences: only single nucleotide substitutions and variations in the number of copies of short reiterations were found in the clinical isolates relative to the laboratory strain. Since the absence of the 134.5 gene homolog was verified in three independent B virus strains, we concluded that it was a genuine feature of the B virus genome.

The protein product of the HSV-1 134.5 gene, infected-cell protein 34.5 (gamma(1)34.5), is a neurovirulence factor with at least two known, distinct functional activities (Chou *et al.*, 1995; Chou *et al.*, 1990; Chou and Roizman, 1992; Chou and Roizman, 1992; Chou and Roizman, 1990; Taha *et al.*, 1989; Thompson *et al.*, 1989). One function, encoded by the carboxyl-terminal domain, negates the antiviral effect of induced protein kinase R by redirecting the host protein phosphatase 1 to dephosphorylate translation initiation factor eIF2, preventing protein synthesis shutoff in infected cells (Chou *et al.*, 1995; Choi *et al.*, 1994; Chou and Roizman, 1992; He *et al.*, 1997; Leib *et al.*, 2000). Another function, mapped to both amino-terminal and carboxyl-terminal domains, somehow enables the virus to replicate in the peripheral and central nervous systems of experimentally infected animals (Chou *et al.*, 1990; Taha *et al.*, 1989). In the absence of 134.5, other genes may supply these functions, given the striking

similarities between the replication characteristics of B virus and the human simplex viruses.

We predict that B virus currently uses compensatory strategies to block host responses to infection, similar to those described for HSV-1. For example, the HSV-1 US11 protein can inhibit protein kinase R activation and compensate for the absence of the ICP34.5 function in deletion mutants if expressed early in infection (Cassady *et al.*, 2002; Cassedy *et al.*, 1998; He *et al.*, 1997; Mohr and Gluzman, 1996). It was proposed that the US11 gene encodes an alternative mechanism to preclude the shutoff of protein synthesis that is currently inactive in HSV-1 (Cassady *et al.*, 2002). In addition, B virus might have evolved unique mechanisms to prevent termination of protein synthesis and elude the DNA replication blocks imposed by neuronal cells. To examine this possibility, a search for B virus-specific genes was performed.

Comparative genomics and codon usage

Two major sequence differences between the B virus and HSV-1 and HSV-2 genomes were detected. The B virus RS region contains an additional 1.5 kb of sequence between the S terminus and the ICP4 gene homolog, while the RL region is shorter in B virus than in the human viruses, with no sequence homology to the HSV ICP0 flanking regions. However, both standard and customized bioinformatics tools have not detected any potential genes in these regions.

A putative two-exon gene (UL53A) was identified by statistical analysis in the B virus UL region on a complementary strand (positions 112186 to 112495 and 112755 to 113344). This putative gene had a codon usage pattern compatible with codons present in

established B virus genes. The UL53A protein is a hydrophilic basic protein (pI 11.9) with an amino acid composition biased to Pro and Arg residues. The first 21 residues were predicted to be a signal peptide by the SignalP program, and the sequence of residues 222 to 240 were predicted to be a transmembrane domain by two out of three programs applied (TMpred and DAS), but the scores were not high enough to classify this protein as a membrane protein. PSI-BLAST analysis detected sequence similarity to domains in three neuronal proteins encoded in mammalian genomes: neural cell adhesion molecule NCAM-180 (GenBank accession number P13595, E-value of 0.16), calcineurin inhibitor cabin 1 (GenBank accession number AAD40846, E-value of 0.014), and brain calcium channel 1A subunit (GenBank accession number AAB64179, E-value of 0.007). However, the similarity was insufficient to confidently characterize the gene at this time, and pending experimental evidence will be critical to determine whether this gene, which is missing in HSV-1 and HSV-2, provides B virus with a unique mechanism to attack neural cells.

The G+C content of the protein-coding regions in B virus, 74.4%, is only slightly lower than the G+C content of the noncoding regions, 75.4%. These data indicate a strong mutation pressure toward G and C nucleotides in the B virus DNA evolution. Obviously, the genetic drift toward the abundance of G's and C's has been compensated for by positive selection for conservative A's and T's in the first (26% A and T) and second (44% A and T) positions of a codon in protein-coding regions and by positive selection for A's and T's in a few evolutionarily conserved regions (promoters, repeats, and other regulatory sites) in noncoding DNA. The G+C content of the third position of codons in the B virus genes is extremely high, 93.1%. This bias, created by mutation

pressure, seems to be a driving force in the formation of the codon usage pattern in B virus.

Genomes of primates tend to have low G+C content, and the logical question arises whether the highly G+C-biased codon usage in the B virus genome matches the proportions of the isoacceptor tRNA in host cells. Unfortunately, no experimental data are currently available regarding the tRNA pool in macaque cells. Still, relative amounts of the isoacceptor tRNAs could be predicted based on the frequencies of codons used in rhesus macaque genes because a strong direct correlation between these parameters has been demonstrated in a number of organisms (Gouy and Gautier, 1982; Ikemura, 1985; Moriyama and Powell, 1997).

A codon usage catalog of rhesus macaque genes was generated based on 288 protein-coding sequences from the GenBank database and compared with the codon usage in B virus genes (Table 3.4). As expected, the overall codon usage patterns of B virus and the macaque hosts differed substantially. However, remarkably, the favored codons of the B virus genes mirrored the most frequently used codons in the rhesus macaque genes (except Ser and Arg, both encoded by six codons). Moreover, these most frequent synonymous codons in the B virus genes were almost the only codons used for each amino acid. This observation correlates with intriguing data from a number of unicellular and multicellular organisms that highly expressed genes utilize preferable codons in each synonymous group, the optimal codons (Duret, 2000; Duret and Mouchiroud, 1999; Gouy and Gautier, 1982; Sharp *et al.*, 1986). Thus, the combined effects of directional mutation pressure and selection for preferential codons brings the B

Table 3.4 - Codon usage in rhesus macaque and B virus genes[a]

Amino acid	Codon	*Macaca mulatta*		B virus	
		No. of occurrences	% of occurrences	No. of occurrences	% of occurrences
His	CAU	972	43	31	3
	CAC	1,284	57	870	97
Tyr	UAU	1,229	43	60	6
	UAC	1,602	57	887	94
Gln	CAA	1,161	30	55	6
	CAG	2,756	70	929	94
Asn	AAU	1,625	46	28	4
	AAC	1,902	54	692	96
Lys	AAA	2,094	45	50	10
	AAG	2,561	55	455	90
Asp	GAU	1,633	44	144	7
	GAC	2,050	56	1,978	93
Glu	GAA	2,141	42	157	7
	GAG	3,018	58	1,986	93
Phe	UUU	1,696	44	268	21
	UUC	2,187	56	1,028	79
Cys	UGU	1,141	47	83	12
	UGC	1,284	53	601	88
Ile	AUU	1,475	32	52	6
	AUC	2,303	50	808	89
	AUA	809	18	47	5
Val	GUU	1,056	19	141	5
	GUC	1,440	25	1,139	41
	GUA	516	09	46	2
	GUG	2,686	47	1,426	52
Pro	CCU	1,366	28	155	4
	CCC	1,476	31	2,014	53
	CCA	1,457	30	120	3
	CCG	518	11	1,527	40
Thr	ACU	1,300	24	31	2
	ACC	2,076	37	992	49
	ACA	1,528	28	58	3
	ACG	577	11	935	46
Gly	GGU	926	17	127	4
	GGC	1,666	30	1,724	54
	GGA	1,641	29	209	6
	GGG	1,352	24	1,180	36
Ala	GCU	1,548	27	144	2
	GCC	2,192	39	3,471	55
	GCA	1,298	23	127	2
	GCG	635	11	2,599	41
Arg	CGU	398	9	82	2
	CGC*	727	17	2,109	**56**
	CGA	368	8	208	5
	CGG	749	17	1,200	31
	AGA*	1,077	**26**	47	1
	AGG	1,014	23	180	5
Leu	UUG	1,109	12	80	2
	UUA	665	7	14	0
	CUU	1,261	14	55	1
	CUC	2,930	21	1,373	36
	CUA	562	6	58	2
	CUG	3,465	40	2,256	59
Ser	UCU	1,428	20	93	4
	UCC*	1,731	**23**	648	29
	UCA	1,122	16	37	2
	UCG*	326	5	778	**35**
	AGU	1,080	15	41	2
	AGC	1,505	21	620	28
Overall G+C			51.03		74.36
First position G or C			52.55		73.73
Second position G or C			42.12		56.22
Third position G or C			58.42		93.11

[a]The sequence of the *M. mulatta* genes were obtained from the GenBank database. The most frequently observed codons are in boldface. Codons for which discrepant results were obtained are marked with an asterisk. Nondegenerate amino acids (Met and Trp) are not included.

virus genes to the favorable position that allows for efficient employment of the host cell translation machinery for expression of viral proteins.

Discussion

The complete genomic sequence of the B virus E2490 laboratory strain demonstrates that the genome organization and gene repertoire are similar but not identical to those of the HSV-1 and HSV-2 genomes. Interestingly, the B virus lacks the gene homolog encoding the neurovirulence factor ICP34.5. We postulate that unless non-orthologous replacement had taken place, the B virus utilizes different mechanisms to block host antivirus responses and facilitate replication in neurons. The enhanced neurovirulence of B virus in human hosts may reflect the inability of human cells to mount efficient antiviral cellular responses against these divergent strategies. The other implication of our findings is that since the B virus does not possess the 134.5 gene homolog, acquisition of this gene by human viruses might be an even more recent evolutionary event than was suggested previously (Cassady *et al.*, 2002; Cassady *et al.*, 1998).

Comparative analysis of the B virus and HSV genomes is essential to understanding the mechanisms of B virus pathogenesis in humans. The complete sequence of the the B virus reference strain also supplies much-needed information to determine the genetic basis of phenotypic and pathogenic differences among the B virus isolates derived from different macaque subspecies. With this information, new antiviral and vaccine strategies can be designed to target critical viral components in order to control this deadly zoonotic agent.

CHAPTER 4

IDENTIFICATION OF PROTEINS ASSOCIATED WITH MURINE CYTOMEGALOVIRUS VIRIONS

Introduction

Proteins associated with the MCMV viral particle were identified by a combined approach using genomic and proteomic methods. Peptides were separated by nano-flow liquid chromatography and analyzed by tandem mass spectrometry (LC-MS/MS) (Kattenhorn *et al.*, 2004). The obtained MS/MS spectra were searched against a database of MCMV ORFs predicted to be protein-coding by a modified version of the gene prediction algorithm, GeneMarkS, which takes into account the heterogeneity of the nucleotide composition as well as other viral specific attributes. The proteins from the capsid, tegument, glycoprotein, replication, and immunomodulatory families were among 38 identified proteins, as well as 20 proteins of unknown function. Observed irregularities in graphs of the coding potential have suggested possible sequencing errors in the 3' proximal ends of m20 and M31. These errors have been experimentally confirmed and corrected by sequencing analysis.

The MCMV sequence exhibits heterogeneity in its nucleotide composition. We found that genes which appeared to be unique to MCMV were clustered in regions with nucleotide composition that differs from one of the regions with genes conserved in other herpesviruses. Thus, our unsupervised method for gene finding in a viral genome would be less accurate, as it would build a model across the whole genome. Using a Bayesian segmentation algorithm developed in the Makeev laboratory, we segmented the MCMV genome into regions of homogeneity and then trained a model for each region separately using our model building procedure. The resulting gene predictions were quite accurate,

with many predictions being verified by the mass spectrometry analysis and/or by similarity searches. In several cases, the predictions helped distinguish real genes from falsely annotated on occassions when two or more overlapping genes were annotated. Previously annotated genes with no experimental support were also revisted.

Statistical gene prediction programs have been used previously for herpesvirus genome annotation. The GeneMark program (Borodovsky and McInich, 1993a), for example, has been used to identify genes in the genomes of bovine herpesvirus 4 (Zimmermann *et al.*, 2001). More recent studies have used information from newly sequenced, closely related CMV genomes (Davison *et al.*, 2003), as well as tools which derive amino acid positional patterns from protein databases to predict protein coding regions in viruses (Bahr and Darai, 2004; Murphey *et al.*, 2003).

In order to help analyze the proteomic data, GeneMark.hmm was used to generate a database of putative MCMV ORFs. This database contains viral ORFs previously annotated (Rawlinson *et al.*, 1996), as well as MCMV genes newly predicted by GeneMark. Searching this database with queries generated by the MS proteomic analysis confirmed a number of annotated MCMV gene products and led to the detection of peptides attributed to novel MCMV genes. In addition, it allowed for detecting sequencing errors in two MCMV virion-associated proteins.

This project (Kattenhorn *et al.*, 2004) was done in collaboration with the Ploegh group at Harvard Medical School. The bench work and mass spectrometry analysis presented in this section was conducted by their group, while the gene predictions, comparative genomics and statistical analyses were done at Georgia Tech.

Methodology

MCMV virion mass spectrometry analysis

The MCMV virions were isolated and analyzed by mass spectrometry. Individual MS/MS spectra were submitted to MASCOT version 1.8 (Matrixscience) (Perkins *et al.*, 1999) and searched either against the NCBI non-redundant database (NCBInr) or against a custom-made database consisting of MCMV gene products predicted the MCMV specific version of GeneMarkS. In addition, two new databases were generated. One contained protein translations of all possible ORFs greater than 20 amino acids. The other included protein translations of all MCMV genomic sequences between two consecutive stop codons. Spectra obtained from the MS/MS analysis were searched by MASCOT against all three databases. These databases allowed for the identification of the MCMV genes by an exhaustive search among all possible virus-derived protein sequences. Proteins were scored using a probabilistic MOWSE algorithm, and the scores were reported in the form of $-10 \cdot \log(P)$, where P is the probability that the observed match is a random event (Perkins *et al.*, 1999). Matches with scores indicating a less-than-5% probability of being a random match were judged as significant. In all searches, oxidation of methionine, carbamidomethylation of cysteine residues, acetylation of N termini, and sodiated glutamic acid and aspartic acid were considered as possible modifications. Individual peptides were ranked as significant if they had a MASCOT MOWSE score of greater than 20 and a minimum of four consecutive b or y ions present in MS/MS spectra. At least two peptides were found for each identified ORF except for ORF105932-106072.

Segmentation and sequence analysis

The sequences of the Smith strain of MCMV and Maastricht strain of rat CMV (RCMV) were obtained from RefSeq (NC_004065 and NC_002512). The analysis of the MCMV genome was performed essentially as described in Chapter 2 (Mills *et al.*, 2003), with the following modifications. The MCMV genome was divided into several regions by using the Bayesian segmentation method (Ramensky *et al.*, 2000), modified to keep the number and the minimal size of segments suitable for subsequent model training (Figure 4.1). The borders between segments were then adjusted to not interrupt any ORF larger than 300 bp.

The individual segments were analyzed by the GeneMarkS program, using either heuristic models defined in the first iteration (for segments shorter than 17 kbp) or models defined at a point of reaching convergence in the self-training process. Since the GeneMark program (Borodovsky and McIninch, 1993), developed prior to GeneMarkS (Besemer *et al*., 2001), has the ability to identify profiles of coding potential that visualize the predicted genes in more detail than GeneMarkS, we also built models for the GeneMark program. These models were built for each genomic segment by using protein-coding and noncoding regions identified by GeneMarkS. GeneMark models were used to generate the GeneMark graphical outputs that were further employed to detect irregularities in the protein-coding potential in the genomic regions around genes of interest. Putative frameshifts identified in this analysis were further considered for additional confirmation through similarity searches in protein databases.

Characterization of the protein product of the previously annotated ORFs which had evidence produced by a statistical analysis was conducted as follows. A PubMed

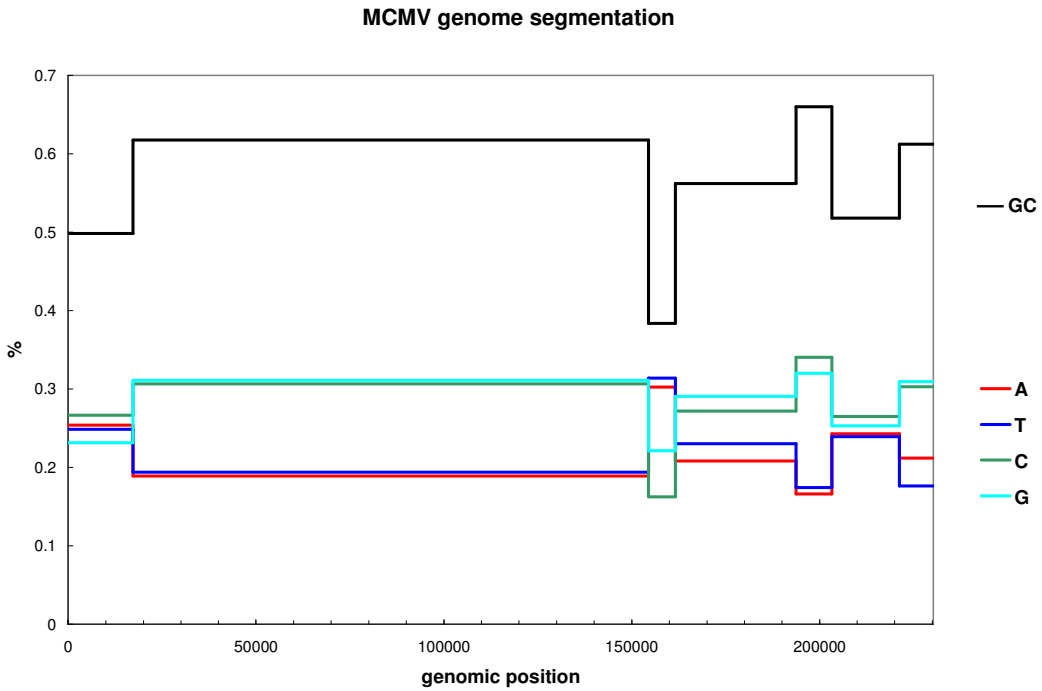

Figure 4.1 - Segmentation of the nucleotide composition of the MCMV complete genome obtained with the use of the Bayesian method (Ramensky *et al.*, 2000)

search was implemented using the ORF name as well as various forms of the virus name (e.g. 'MCMV', 'Mouse Cytomegalovirus', 'Murine Cytomegalovirus', etc). This search was done to gather any positive or negative experimental evidence that could exist for the particular ORF. Next, both BLAST and RPS-BLAST (Altschul *et al.*, 1997) searches were conducted for the translated protein to determine whether any similar proteins had already been characterized. In addition, the genomic context of the ORF was checked in RCMV for positional orthologs with characterized function.

Results

Analysis of MCMV virions

MCMV virions were isolated and analyzed with mass spectrometry. Due to the incompleteness of current MCMV annotations in public protein databases, it was necessary to construct a unique database of predicted MCMV gene translations against which to search the peptides obtained by MS/MS. The new database was created based on the GeneMarkS program predictions. In an attempt to prevent search bias due to the small database size, searches were also conducted against a merged database of SwissProt and the newly assembled MCMV entries. The large size of the latter database, however, greatly increased the MASCOT MOWSE cutoff score and decreased the sensitivity of this approach. Using these newly constructed databases, 19 viral proteins were identified by the MS/MS analysis of the in-gel digested polypeptides (Figure 4.2B).

In order to identify the virion-associated components of lower abundance, direct in-solution digestion of isolated virions was carried out in parallel. This approach revealed a total of 58 virus-derived gene products associated with the virion (Table 4.1).

All four annotated MCMV capsid proteins, m48.2, M85, M86, and M46, were found associated with MCMV virions by the in-solution digestion approach. In contrast, only two capsid proteins, M85 and M86, were recovered from in-gel digestion. M46 is the homologue of UL46, the minor capsid protein of HCMV. UL46 aggregates upon heating and is unable to enter the resolving gel during SDS-PAGE (Gibson *et al.*, 1996). This characteristic of the protein appears to apply to M46 as well. m48.2 has a predicted molecular mass of only 9.8 kDa. Therefore, it was not resolved by the 10% acrylamide matrix used in this study.

The MCMV glycoproteins gB, gM, and gH (M55, M100, and M75, respectively) were identified by the in-solution digestion approach. In addition, m74, the positional homologue of HCMV gO, was also identified. In-gel analysis identified only M55 and M100. gL, previously identified as associated with purified virions (Xu *et al.*, 1994), was not found by either in-gel or in-solution analysis.

In-solution digestion identified nine proteins homologous to HCMV tegument proteins: lower and upper matrix phosphoproteins (M83 and M82), large tegument protein (M48), pp150 (M32), M25, M47, M94, M99, and M51 (Table 4.1. In-gel analysis detected seven of these proteins (Figure 4.2B). Although M78 has been reported to be an MCMV tegument protein (Oliveira and Shenk, 2001), it was not detected by either in-solution digestion analysis or in-gel digestion analysis.

Eleven loci encoding *trans*-acting factors have been identified as being required for transient complementation of HCMV oriLyt-dependent DNA replication (Pari and Anders, 1993). Six of these proteins comprise the structure required to initiate and perform DNA synthesis. Five of these homologues in MCMV, DNA polymerase (M54),

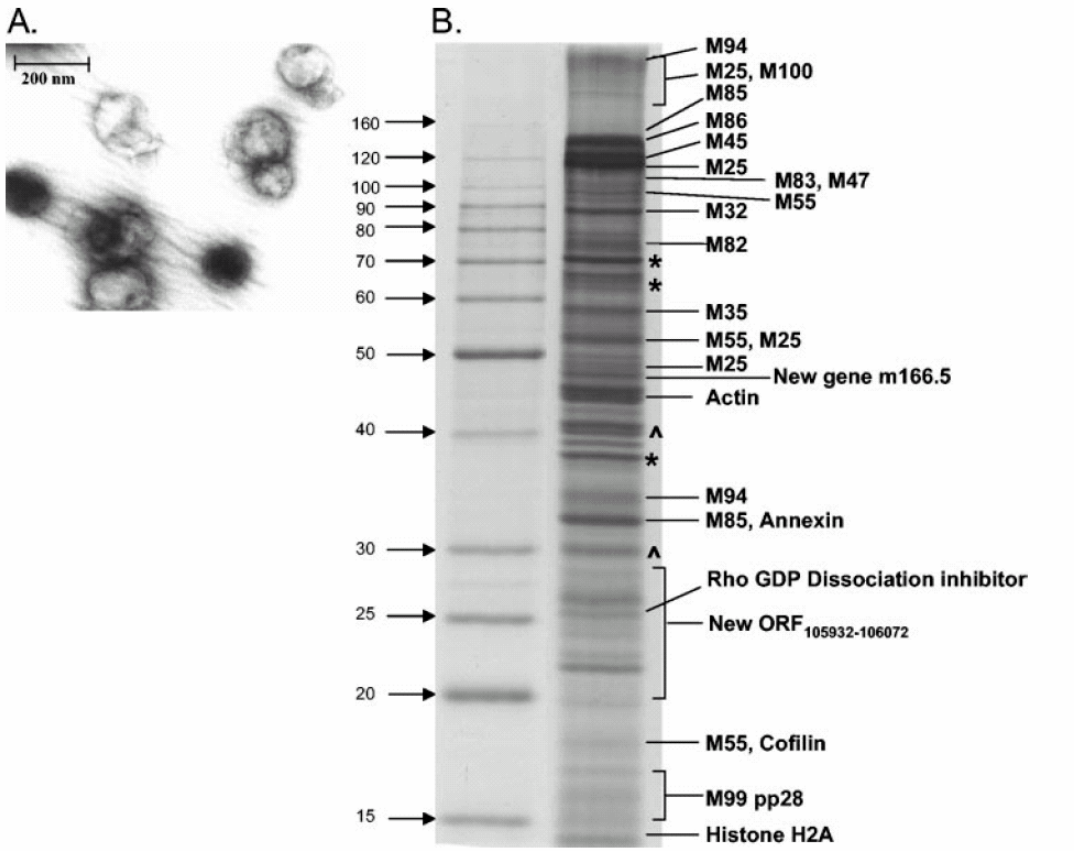

Figure 4.2 - Characterization of isolated MCMV virions. (A) Electron microscopy picture of purified MCMV particles. (B) Analysis of virus particles isolated from supernatant of infected cells. MCMV-encoded polypeptides and host cell proteins are indicated, as identified using the compiled database of predicted proteins.

Table 4.1 - MCMV ORFs associated with MCMV virions identified by MS/MS

ORF	Strand	Position		HCMV[e]	Comment (reference)[a]	Mass (kDa)		% Coverage[f]	No. of spectra[g]
		From	To			Predicted	Observed		
m02		999	1979		Glycoprotein family m02 (36)	36.6		39	10
m18	C	17071	20193		Highly antigenic early gene (23)	108.5		14	6
m20[b]	C	20581	23044			89.4		6	2
M25		26015	28813	UL25	Tegument protein (60)	130, 105, 95	48	54	91
m25.2	C	30245	31657	US22	US22 homologue (30)	51.8		44	9
M28	C	34486	35778	UL28		47.3		15	4
M31[c]		37278	39068	UL31		66.1		18	5
M32	C	39280	41436	UL32	(HCMV: pp150) (25)	78.6	90	37	20
M35		47465	45909	UL35	UL25 family member, virulence factor (52)	58.1	58	41	8
m39	C	52484	53200			25.6		59	4
M43	C	55351	57144	UL43	Immunoregulatory gene (49)	67.0		55	11
M44	C	57885	59120	UL44	DNA binding phosphoprotein (27)	44.6		22	3
M45	C	59515	63038	UL45	Ribonucleotide reductase homolog, anti-apoptotic (11)	120	120	34	12
M46	C	63040	63924	UL46	(HCMV: minor capsid binding protein) (20)	33.2		36	8
M47		63923	67045	UL47	(HCMV: tegument protein) (4)	118.1	115	9	3
M48		67042	73491	UL48	Large tegument protein (45)	238.5		22	18
m48.2	C	73571	73867	UL48.5	Smallest capsid protein (8)	9.8	87	9	
M51	C	76515	77216	UL51		25.1	27	2	
M54	C	79701	82994	UL54	DNA polymerase (45)	123.8		17	8
M55	C	83003	85816	UL55	Glycoprotein B (44)	130, 105, 52	99, 52, 18	41	20
M56	C	85716	88112	UL56	(HCMV: terminase subunit, tegument protein) (7)	89.0		25	8
M57	C	88319	91894	UL57	(HCMV: Single-stranded DNA binding protein) (39)	131.4		10	6
M69	C	96193	98721	UL69	(HCMV: tegument protein) (58)	93.0		28	11
M70	C	99010	101904	UL70	(HCMV: helicase-primase subunit) (39)	109.8		11	5
M71		101903	102802	UL71		32.9		65	6
M72	C	103031	104236	UL72	dUTPase (31)	45.0		13	2
M74	C	104496	105812	UL74		49.1		27	6
M75	C	106110	108287	UL75	Glycoprotein H (61)	81.3		32	8
M77		108931	110817	UL77	(HCMV: pyruvoyl decarboxylase) (63)	68.6		5	2
M80		113414	115507	UL80	Assembly protein-protease (28)	74		7	4
M82	C	115711	117507	UL82	Upper matrix phosphoprotein, pp71 (16)	67.4	75	32	11
M83	C	117614	120043	UL83	Lower matrix phosphoprotein, pp65 (16)	90.9	110	20	10
M85	C	122192	123124	UL85	Minor capsid protein (only found in gel) (3)	34.6	34, 140	52	15
M86	C	123199	127260	UL86	(HCMV: major capsid protein) (12)	151.4	150	42	41
M87		127383	130163	UL87		102.4		21	6
M88		130243	131523	UL88	(HCMV: virion protein) (3)	47.3		21	4
m90	C	132920	133876			35.8		17	4
M94		136234	137271	UL94	(HCMV: virion-associated protein) (57)	37.7	>160, 35	55	23
M95		138282	139535	UL95		45.8		20	5
M97		140141	142072	UL97	Phosphotransferase (46)	71.1		47	10
M98		142101	143769	UL98	(HCMV: alkaline nuclease) (48)	62.0		23	4
M99		143723	144061	UL99	Virion-associated phosphoprotein (15)	11.9	15–17	67	8
M100	C	144296	145411	UL100	Glycoprotein M (47)	42.3	>160	50	15
M102		145596	148034	UL102	(HCMV: helicase-primase subunit) (39)	91.0		24	8
M104	C	149113	151227	UL104	Structural protein (50)	80.6		31	12
M105		151028	153874	UL105	(HCMV: helicase-primase subunit) (39)	106.4		40	11
m107		161983	162678			24.6		29	4
M116	C	167205	169142	UL116		66.1		6	2
m117.1		169541	170956			45.3		7	2
M121	C	175679	177775	UL121		73.2		28	6
m147	C	206862	207299			16.9		32	6
m150	C	208789	207623		Member of MGP[h] family m145 (45)	42.8		28	5
m151	C	208814	209983		Member of MGP family m145 (45)	42.4		28	3
m163	C	221875	222645			19.1		19	3
m165	C	223280	224278			35.8		10	2
m166.5[d]	C	225441	226898		New gene	48	42, 48	31	10

[a]Comments in parentheses preceded by "HCMV:" indicate a gene function annotated solely from the homologue in HCMV.
[b]Different 5' extension based on a frameshift sequencing error at position 20958 in the original sequence.
[c]Extended C terminus based on insertion of a G at position 38803.
[d]New gene located between m166 and m167.
[e]HCMV homologue.
[f]Percent coverage of protein by polypeptides detected by MS.
[g]Number of polypeptides detected from this ORF by MS.
[h]MGP, potential membrane glycoprotein.

polymerase accessory protein (M44), major DNA binding protein (M57), and all proposed subunits of the helicase-primase complex (M70, M102, and M105) were found with high degrees of confidence by the in-solution digestion of the virion preparations (Table 4.1).

The virion preparation contained a number of proteins suggested to interact with the immune system, including M43, a protein thought to influence T helper cell responses *in vivo* (Singh *et al.*, 2003), and M45, a protein thought to have multiple functions during viral infection (Brune *et al.*, 2001; Patrone *et al.*, 2003). In-gel digestion analysis identified one of these proteins.

In addition to known tegument and capsid proteins, a number of other viral proteins were identified in association with purified virions. The assembly protein M80 and the DNA packaging protein M56 were both found. In addition M98, the viral exonuclease, and M97, the viral protein kinase, were present. Lastly, M72, the dUTPase, was detected albeit with a slightly lower degree of confidence. This suggests it may be present in low copy numbers per virion.

Gene and protein characterization

The complete genome of MCMV was analyzed as described and compared to the original annotation (Table 4.2). In addition to 131 genes previously known, 12 new genes were identified and 31 genes originally annotated were not confirmed. Also, 12 predictions overlapped the exons of the genes known to contain introns.

Several proteins of unknown function, not previously described to be present in the virion, were identified by the MS/MS analysis of the in-gel digested polypeptides.

Table 4.2 - Results of the whole genome analysis of MCMV using segmentation coupled with the gene finding procedure

Category	Number
Exact match between prediction and annotation	125
Predicted gene differs in start location with annotated one	6
Predicted gene overlaps with an intron containing annotated gene	12
Original annotated gene was not predicted	31
Newly predicted genes	12

These included m18, m25.2, M28, M31, M35, m39, M71, M87, M88, m90, M95, m107, M121, m150, m151, m163, and m165 (Table 4.1). These proteins were not detected by analysis of in-gel fragments, indicating that they are probably present at low copy numbers. Identification of these proteins as virion associated may provide a direction to further in vivo and in vitro characterizations of their function.

MS/MS analysis identified two peptides with significant MOWSE scores originating from ORF c20579-21313 (Figure 4.3A to C). The peptides appeared to cluster in the beginning of the reading frame and terminate where m20 begins in reading frame 2. GeneMarkS predicted a high coding potential for this region (Figure 4.3C). This suggested that the identified peptides belong to the extension of m20, shifted to a different reading frame by a sequencing error in the reported sequence. Sequence analysis of region 20,600 to 21,200 confirmed that the nucleotide G at position 20,958 in the then current sequence was incorrect (Figure 4.3D). This changes the C-terminal end of m20, extending the ORF sequence from the originally annotated position 20,805 to the newly predicted position 20,579.

Similar to the case of m20, a small ORF located C terminally to the M31 gene, but in a different reading frame, was predicted to have high coding potential (Figure. 4.4A). Resequencing of the M31 gene region revealed an extra G nucleotide at position 38,803 (Figure 4.4B). This additional nucleotide shifts the reading frame of the second half of M31 to encompass this newly identified coding region and also restores full-length homology of M31 to the R31 protein of RCMV (Figure 4.4C).

MASCOT analysis of our MS/MS data using a database of translations of all possible ORFs or all possible regions between stop codons of MCMV led to

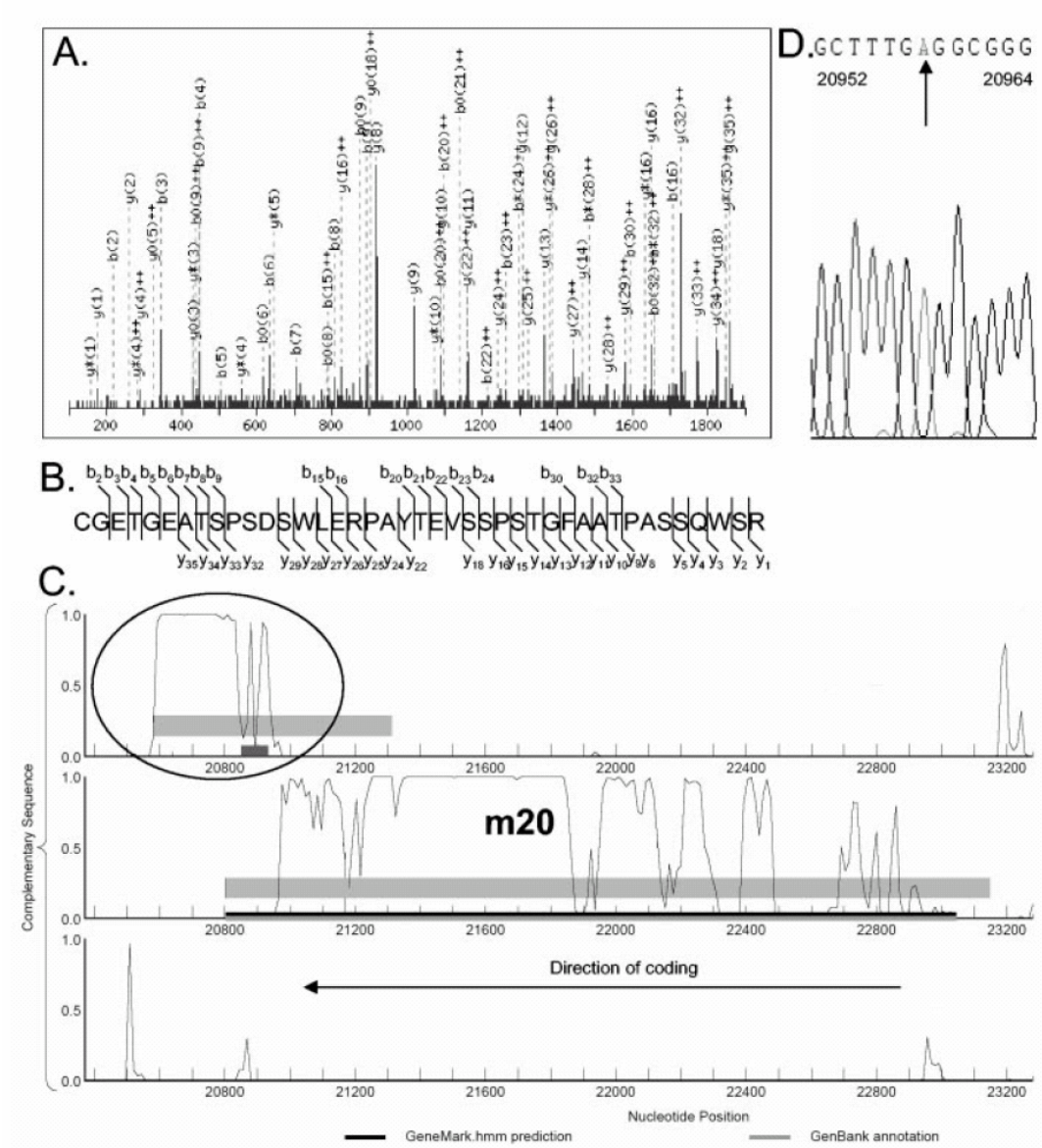

Figure 4.3 - Evidence for a 3' extension of the MCMV m20 gene. (A) MS/MS spectrum of the 1,449.75 $(M+H)^{3+}$. precursor ion of the tryptic peptide fragment CGETGEATSPSDSWLERPAYTEVSSPSTGFAATPASSQWSR as interpreted by MASCOT with a score 45. (B) The matching b and y ions of the peptide found by MS/MS analysis. (C) Putative coding region at position 20579 to 21313 overlapping the 5' region of the m20 gene. The three reading frames of the complementary sequence are displayed. The location of the tryptic peptide identified from this ORF, corresponding to the MS/MS spectrum in panel A is indicated as a black box. Thin grey bars indicate genes annotated by Rawlinson et al. Black bars represent ORFs predicted as protein coding by GeneMarkS, and grey areas indicate regions of interest with moderate coding potential generated by GeneMark. (D) DNA sequencing analysis revealed an incorrect insertion (G) at position 20958 of the MCMV genome (Smith strain NCBI 004065). Removal of this G extended the C terminus of m20 from position 20805 to position 20579.

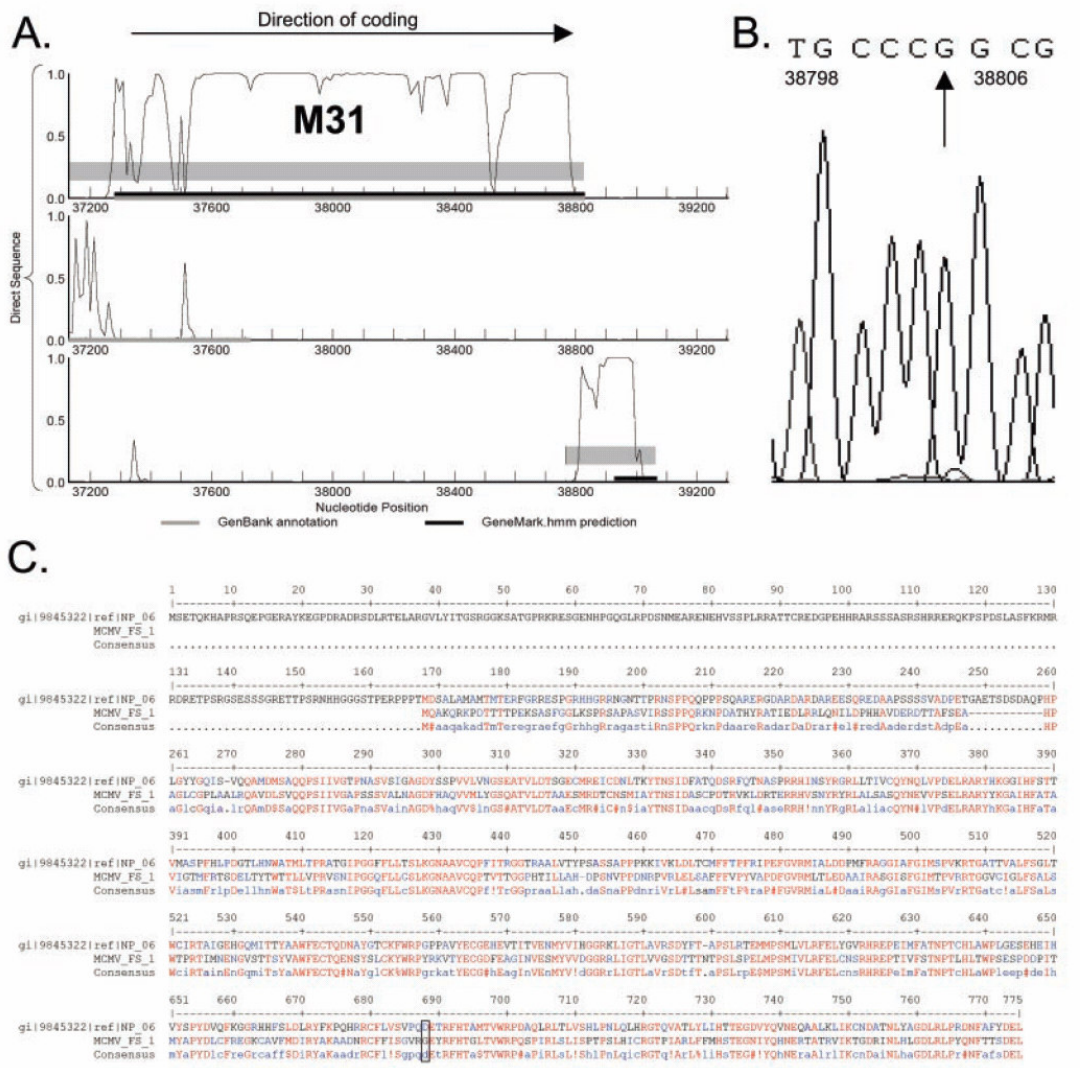

Figure 4.4 - Identification of a frameshift in the MCMV M31 gene. (A) A region with high coding potential was detected at the C-terminal end of the M31 gene in reading frame 3. The three reading frames of the direct sequence are displayed. (B) DNA sequencing analysis determined a missing G at position 38803. (C) Insertion of G38803 restored full-length homology of M31 to RCMV protein R31. The box indicates the amino acid which corresponds to the corrected codon from the insertion.

identification of peptides derived from two previously unannotated ORFs. In-solution digestion analysis and SDS-PAGE analysis of virus material identified peptides from MCMV ORFc 225441-226898. This ORF is located between m166 and m167 on the complementary strand of MCMV (Figure 4.5) and specifies an ORF of 446 amino acids, hereafter referred to as m166.5 (Figure 4.5B). It overlaps significantly with m167 and was marginally predicted by GeneMark (Figure 4.5C). Analysis of virus material by SDS-PAGE revealed a polypeptide migrating at 4k. In-solution digestion and SDS-PAGE analysis identified 10 peptides from this region, resulting in 31% sequence coverage (Figure 4.4B and C). Peptides from m166.5 were detected in two out of three in-solution virus preparations.

To confirm expression of the newly predicted m166.5 gene, NIH 3T3 cells were infected with recombinant MCMV-m166.5-HA possessing a C-terminal fusion of the HA tag to the m166.5 ORF. As control, the recombinant virus MCMV-m166-HA with an HA fusion to the m166 ORF was used. Total cell lysates were harvested 5 and 24 h postinfection, and immunoblot analysis was performed. As shown in Fig. 4.5D, a specific band for the HA-tagged m166.5 ORF with a size slightly larger than 45 kDa could be detected at 5 and 24 h postinfection. These data confirmed that a protein is expressed from ORF m166.5 during MCMV infection.

A peptide with a significant MOWSE score was also identified from the region of ORF105932-106072 by in-solution digest analysis. This ORF codes for a possible protein of 44 amino acids is located in an unannotated region of the MCMV genome and is predicted by GeneMark.hmm to have high coding potential (Figure 4.6). The virus-

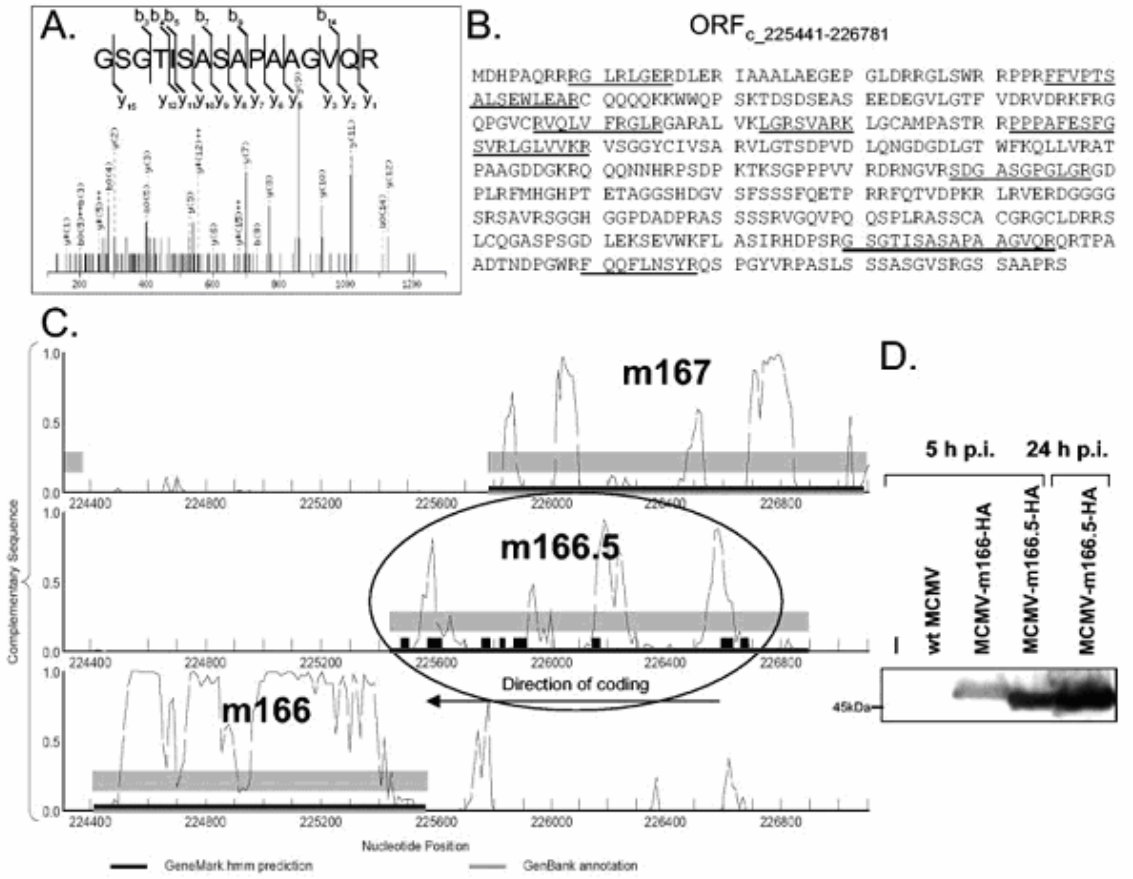

Figure 4.5 - Expression of m166.5 confirmed by the MS/MS and Western blot analysis. (A) MS/MS spectrum of precursor ion 715.26 (uppercase 2+) was interpreted as peptide GSGTISASAPAAGVQR by MASCOT, with a score of 44. Detected b and y ions are indicated. (B) Sequence of m166.5. Tryptic peptides detected by MS/MS are underlined. (C) GeneMark probability plot of the m166 MCMV gene region. The three reading frames of the complementary sequence are displayed. m166.5 partially overlaps with m167, and similar probabilities of coding potential are seen for both reading frames. Peptides found by MS/MS analysis are indicated as black boxes. Thin grey bars represent genes annotated by Rawlinson et al., black bars indicate genes predicted by GeneMarkS, and grey areas indicate regions of interest with moderate coding potential. (D) Western blot analysis of NIH 3T3 endothelial cells infected with MCMV containing an HA-tagged m166.5 gene. Infected cells were harvested after the indicated times, lysed, and analyzed by SDS-PAGE and immunoblotting with an anti-HA antibody.

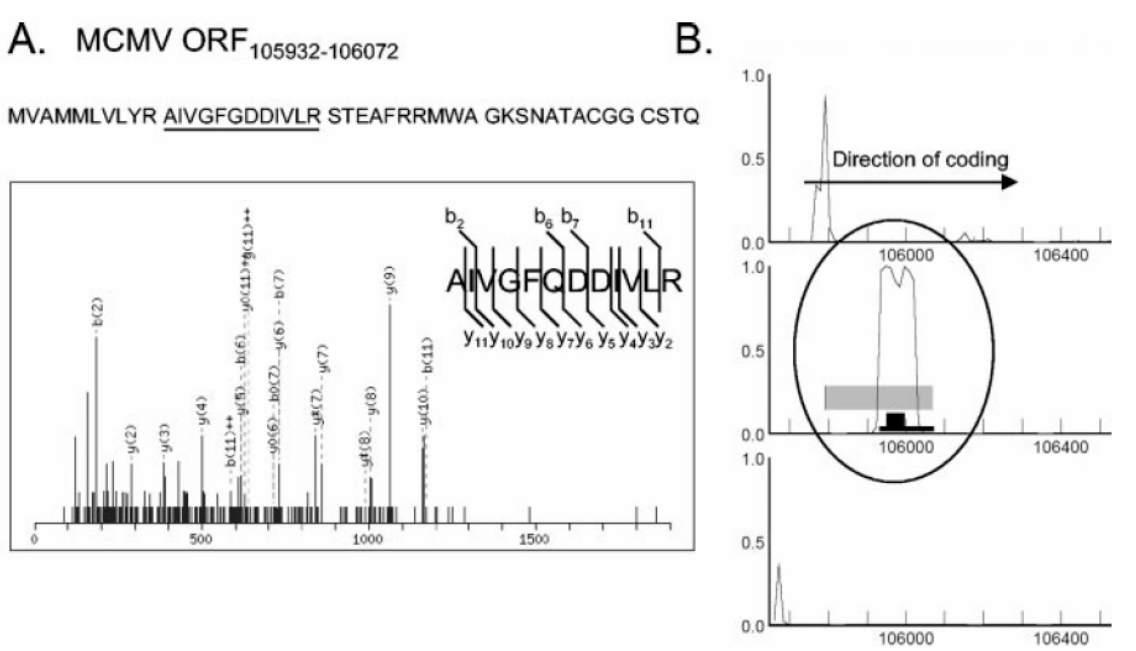

Figure 4.6 - Evidence for expression of the predicted ORF105932–10672. (A) The MS/MS spectrum of the precursor ion 673.21 (uppercase 2+) corresponds to the tryptic fragment AIVGFQDDIVLR with a MASCOT score of 48. Matching b and y ions are indicated. This virus-derived peptide was detected in the virion preparation analyzed by SDS-PAGE and isolated from the 20- to 29-kDa range. The tryptic peptide fragment detected within the sequence of ORF105932–106072 by MS/MS is underlined. (B) Gene with high coding potential predicted by GeneMark.hmm. The three reading frames of the direct sequence are displayed. A black box indicates the tryptic fragment identified by MS/MS. The predicted gene and regions with moderate coding potential are indicated as black and gray bars, respectively.

Table 4.3 - Cellular proteins associated with MCMV particles

Cellular protein	SwissProt ID	Other matches[a]	Mass (kDa)		Comment (reference)	MS spectra[b]	% Coverage[c]
			Predicted	Observed			
γ-Actin, cytoplasmic	gi:6752954	β-actin, α-actin 1	42	45	Associated with HCMV preparations (3)	14	42
Annexin I (lipocortin I)	gi:1351942	Annexin V	35.7	33	Associated with HCMV preparations (59)	3	11
Cofilin	gi:116849		18.5	18	Associated with HIV particles (38a)	2	16
EF-1 α (EF-Tu)	gi:1169475		50			5	22
Glyceraldehyde-3-phosphate dehydrogenase	gi:120702		35.7			1	6
Histone H2A	gi:121961	Histone H2B	14	14		6	20
Rho GDP dissociation inhibitor	gi:21759130		23.3	26		2	7

[a]Other cellular proteins matching the same set of peptides.
[b]Number of polypeptides detected from this protein by MS.
[c]Percent coverage of protein by polypeptides detected by MS.

derived peptide was also isolated from the 20- to 29-kDa range of our SDS-PAGE gel during in-gel analysis (Figure 4.2B).

Identification of peptides from a number of cellular proteins was accomplished by conducting MASCOT searches using the NCBInr database. The results from in-gel analysis mirrored those obtained from in-solution digestion, with actin being the predominant cellular protein. In addition, annexin, cofilin, histone H2A, elongation factor 1, glyceraldehyde-3-phosphate dehydrogenase, and rho GDP dissociation inhibitor were present and detectable by in-solution and in-gel digestion (Figure 4.2B and Table 4.3).

Previously annotated but not predicted genes were analyzed further. For 17 of them we did not find any data to support the annotation (Table 4.4). Several ORFs were characterized as highly hypothetical genes due to an almost complete overlap with a gene supported by computational and sometimes experimental evidence, for example m30, m69.1, and m48.1 (Fig. 4.7). In the case of m29, researchers found an additional guanine residue at position 36198, shortening the ORF from 981 to 729 (Habib et al., 2003). This smaller ORF still lacks experimental support and overlaps significantly with well characterized genes. In another example, the m130 ORF overlapping the MCMV chemokine homolog m131/129 transcript (Fleming et al., 1999) does not possess any evidence for being a functional gene.

Discussion

We identified the majority of CMV virion protein constituents previously reported in the literature. The in-solution digestion analysis confirmed the presence of all four capsid proteins, M85, M86, m48.2, and M46. In addition, all but one of the reported

Table 4.4 - Highly hypothetical genes, as inferred from statistical analysis and available data.

MCMV ORF	Length (aa)	Description
m01	118	No statistical or experimental support
m19	147	No statistical or experimental support
m22	104	No statistical or experimental support
m29	327 / 243	Additional guanine residue at position 36198(Habib) shortens ORF by 84 amino acids. ORF overlaps both M28 and m29.1. No statistical or experimental support
m30	281	ORF overlaps M31. No statistical or experimental support
m48.1	103	ORF overlaps m48.2. No statistical or experimental support
m58	234	ORF positional homolog of r58, similarity only in region which oveoverlaps M57 and R57 respectively. No statistical or experimental support
m69.1	119	ORF overlaps M69. No statistical or experimental support
m106	147	ORF located near A-T rich region. No statistical or experimental support
m107	231	ORF overlaps m108. No statistical or experimental support
m119.5	111	ORF overlaps m119.4 and m120. No statistical or experimental support
m124.1	135	ORF overlaps m124. No statistical or experimental support
m126	91	No statistical or experimental support
m130	157	ORF overlaps m129 and m131(Fleming). No statistical or experimental support
m148	145	ORF overlaps m147. No statistical or experimental support
m149	229	ORF overlaps M150 and putative gene at 207314..207502. No statistical or experimental support
m156	147	No statistical or experimental support

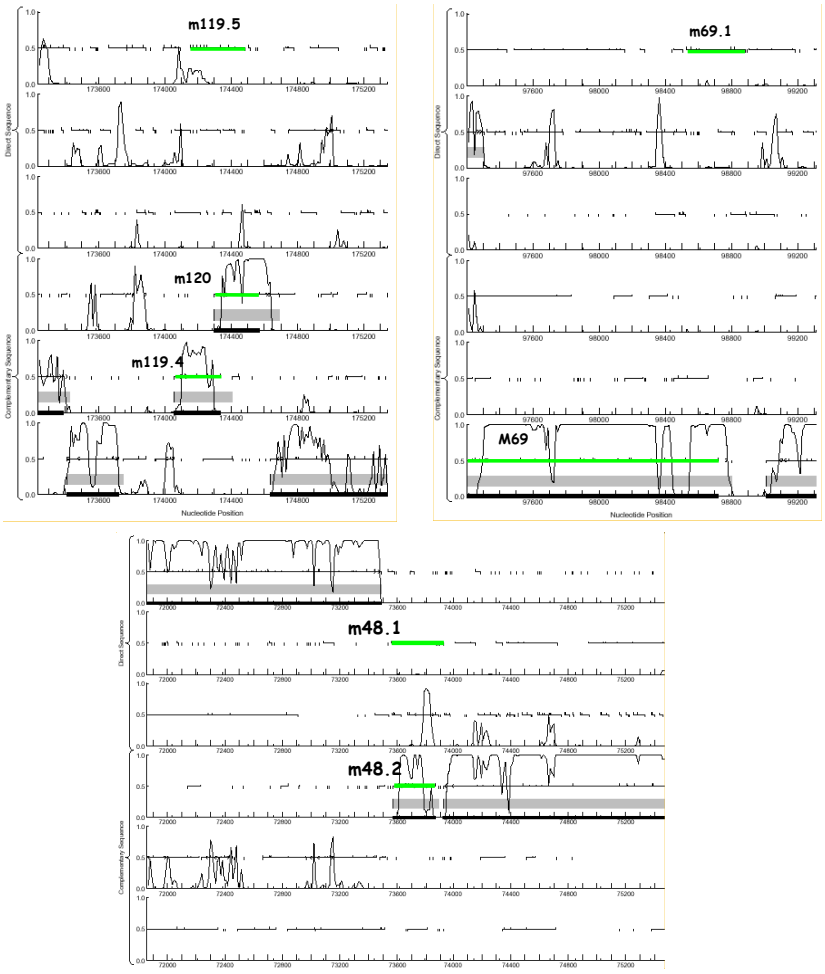

Figure 4.7 - Examples of ambiguous annotated ORFs in MCMV which can be distinguished by statistical analysis; m119.5 (a), m69.1 (b), and m48.1(c) are unlikely to be real.

MCMV tegument proteins were unambiguously detected by this method. M78, a G-protein-coupled, seven-transmembrane receptor has previously been described as a virion component (Oliveira and Shenk, 2001). We only detected a single peptide derived from M78 in one out of four virion preparations. It did not, therefore, meet our threshold criteria to be considered significant. The reason for this discrepancy is not clear, but the transmembrane structure of this protein may account for its inaccessibility to tryptic digestion.

A common approach to the herpesvirus genome annotation is to assign as a protein coding region any ORF longer than 300nt with less than 60% overlap with an adjacent ORF (Rawlinson *et al.*, 1996; Vink *et al.*, 2000). These rules can cause substantial overannotation, especially in genomes that have high G+C content, such as MCMV. The prediction algorithm used in this study, an MCMV-modified version of GeneMarkS, takes into account features of viral genome organization, including small genome size, presence of overlapping genes and repeats, as well as its circular shape.

Computer analysis of the MCMV genome using GeneMarkS, as described above, led to the discovery of 12 putative novel genes. The complementary proteomic analysis suggested the existence of two of these genes as being present in the virion. Expression of one of these genes, m166.5, during viral infection was confirmed by generating recombinant MCMV expressing an HA-tagged form of this protein. This recombinant virus will be a useful instrument for further exploring the biological function of this viral gene.

The second ORF identified, ORF105932-106072, codes for a possible translation product of 44 amino acids in an unannotated region of the MCMV genome. BLAST

search using the NCBI server and the SwissProt database revealed sequence identity to a fragment of a putative spliced M73 transcript described as an unpublished finding by Dallas et al. (gi:762817). This suggests that this fragment may be part of a larger, spliced gene. Consistent with this hypothesis, the peptide derived from this region was detected by SDS-PAGE analysis in the range of 20 to 29 kDa. The full sequence of this gene remains unknown, as peptides were not detected that corresponded to the rest of the M73 putative spliced sequence.

The graphical output generated by GeneMark allowed for visual examination of irregularities in the protein-coding potential of the genomic regions of interest. This examination led to the identification of sequencing errors in the m20 and M31 genes. Restoration of proper coding sequence led to establishing the full-length homology of the M31 gene product to its R31 counterpart in RCMV (Figure 4.4C).

The inability to detect other newly predicted genes may be due to the fact that, even if these ORFs are expressed, the products encoded may not be structural proteins, or they may be of too low abundance to be detected. Further experiments are required to address these possibilities.

The combined analysis described above represents an improved strategy for determining the protein composition of virus particles. The standard method of virion analysis by SDS-PAGE was compared to in-solution digestion and MS/MS analysis of viral proteins. In-solution digestion followed by MS/MS greatly improved sensitivity, increasing the number of identified proteins from 19 (in-gel analysis) to 58 (in-solution analysis). Although the functional consequences of the presence of the newly identified proteins in the virion for viral life cycle are not entirely clear at this time, the data

indicate that the particle composition of MCMV is more complex than previously thought.

REFERENCES

Afonso, C.L., Tulman,E.R., Lu,Z., Zsak,L., Rock,D.L, and Kutish,G.F. (2001). The genome of turkey herpesvirus. *J. Virol.*, **75**, pp 971–978.

Atalay,R., Zimmermann,A., Wagner,M., Borst,E., Benz,C., Messerle,M., and Hengel,H. (2002) Identification and expression of human cytomegalovirus transcription units coding for two distinct Fcγ receptor homologs. *J. Virol.*, **76**, pp. 8596–8608.

Bahr,U., and Darai,G. (2001) Analysis and characterization of the complete genome of tupaia (tree shrew) herpesvirus. *J. Virol.*, **75**, pp 4854–4870.

Bahr,U., and Darai,G. (2004) Re-evaluation and in silico annotation of the Tupaia herpesvirus proteins. *Virus Genes*, **28**, pp. 99–120.

Baldick,C.J.,Jr., and Shenk,T. (1996) Proteins associated with purified human cytomegalovirus particles. *J. Virol.*, **70**, pp. 6097–6105.

Batzoglou S., L. Pachter, J. P. Mesirov, B. Berger, E. S. Lander (2000). "Human and mouse gene structure: comparative analysis and application to exon prediction". *Genome Res*., **10**(7): 950-958.

Belanger,A.E., Hendrix,R.W., and Hatfull,G.F. (1998) Genome structure of mycobacteriophage D29: implications for phage evolution. *J. Mol. Biol.*, **279**, pp.143–164.

Bennett,A.M., Harrington,L., and Kelly,D.C. (1992) Nucleotide sequence analysis of genes encoding glycoproteins D and J in simian herpes B virus. *J. Gen. Virol.*, **73**, pp 2963–2967.

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., and Wheeler,D.L. (2005) GenBank. *Nucleic Acids Res.,* **33**, D34-D38

Besemer,J. and Borodovsky,M. (1999) Heuristic approach to deriving models for gene Finding. *Nucleic Acids Res.*, **27**, pp. 3911-3920.

Besemer,J., Lomsadze,A., and Borodovsky,M. (2001) GeneMarkS: a selftraining method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, pp.2607–2618.

Birney, E. and R. Durbin (1997). "Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison". *Proc Int Conf Intell Syst Mol Biol*., **5**:56-64.

Birney, E. and R. Durbin (2000). "Using GeneWise in the Drosophila annotation Experiment". *Genome Res*., **10**(4): 547-548.

Blattner,F.R., Plunkett III,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., et al. (1997) The complete genome sequence of Escherichia coli K-12. *Science,* **277**(5331), pp. 1453-1474.

Borodovsky, M., Sprizhitsky, Yu. A., Golovanov, E. I., Alexandrov, A. A. (1986a) Statistical features in the Escherichia coli genome functional primary structure. III. Computer recognition of protein coding regions. *Molecular Biology* **20**, 1144-1150

Borodovsky, M., and McInich,J.. (1993a) GeneMark: parallel gene recognition for both DNA strands. *Comput. Chem*., **17**, pp. 123–133.

Borodovsky,M., and McIninch,J. (1993b) Recognition of genes in DNA sequence with ambiguities. *Biosystems*, **30**, pp 161–171.

Borodovsky,M., Rudd,K.E. and Koonin,E.V. (1994) Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res.*, **22**, pp. 4756-4747.

Bortz,E., Whitelegge,J.P., Jia,Q., Zhou,Z.H., Stewart,J.P., Wu,T.T., and Sun,R. (2003) Identification of proteins associated with murine gammaherpesvirus 68 virions. *J. Virol.*, **77**, pp. 13425–13432.

Bresnahan,W.A., and Shenk,T. (2000) A subset of viral transcripts packaged within human cytomegalovirus particles. *Science*, **288**, pp. 2373–2376.

Brune,W., Menard,C., Heesemann,J., and Koszinowski,U.H. (2001) A ribonucleotide reductase homolog of cytomegalovirus and endothelial cell tropism. *Science*, **291**, pp. 303–305.

Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D. (1996) Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. *Science*, **273**(5278), pp. 1058-73.

Burge,C., Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, pp. 78-94

Burland,V., Shao,Y., Perna,N.T., Plunkett,G., Sofia,H.J. and Blattner,F.R. (1998) The complete DNA sequence and analysis of the large virulence plasmid of Escherichia coli O157:H7. *Nucleic Acids Res.*, **26**, pp. 4196-4202.

Burset, M., R. Guigo (1996). "Evaluation of gene structure prediction programs". *Genomics*, **34**(3): 353-367.

Cassady, K.A., Gross,M., and Roizman,B. (1998) The herpes simplex virus US11 protein effectively compensates for the γ134.5 gene if present before activation of protein kinase R by precluding its phosphorylation and that of the alpha subunit of eukaryotic translation initiation factor 2. *J. Virol.*, **72**, pp. 8620–8626.

Cassady,K.A., Gross,M. and Roizman,B. (1998) The second-site mutation in herpes simplex virus recombinants lacking the γ134.5 genes precludes shutoff of protein synthesis by blocking the phosphorylation of eIF-2α. *J. Virol.*, **72**, pp. 7005–7011.

Cassady,K.A., Gross,M., Gillespie,G.Y. and Roizman,B. (2002) Second-site mutation outside of the US10–12 domain of γ134.5 herpes simplex virus 1 recombinant blocks the shutoff of protein synthesis induced by activated protein kinase R and partially restores neurovirulence. *J. Virol.*, **76**, pp. 942–949.

Cawley,S., Pachter,L., and Alexandersson,M. (2003) SLAM web server for comparative gene finding and alignment. *Nucl. Acids. Res.*, **31**, pp. 3507-3509

C. elegans Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, **282**(5396), pp. 2012-8

Chang,Y.E., Menotti,L., Filatov,F., Campadelli-Fiume,G., and Roizman,B.. (1998) UL27.5 is a novel γ2 gene antisense to the herpes simplex virus 1 gene encoding glycoprotein B. *J. Virol.*, **72**, pp 6056–6064.

Chelius,D., Huhmer,A.F., Shieh,C.H., Lehmberg,E., Traina,J.A., Slattery,T.K., and Pungor Jr.,E. (2002). Analysis of the adenovirus type 5 proteome by liquid chromatography and tandem mass spectrometry methods. *J. Proteome Res.*, **1**, pp. 501–513.

Cherry,J.M., Ball,C., Weng,S.., Juvik,G., Schmidt,R., Adler,C., Dunn,B., Dwight,S., Riles,L., Mortimer,R.K., Botstein,D. (1997) Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, **387**, pp. 67-73

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature,* **437**, pp.69–87

Chong,K.T., and Mims,C.A., (1981) Murine cytomegalovirus particle types in relation to sources of virus and pathogenicity. *J. Gen. Virol.*, **57**, pp. 415–419.

Chou,J., and Roizman,B. (1990) The herpes simplex virus 1 gene for ICP34.5, which maps in inverted repeats, is conserved in several limitedpassage isolates but not in strain 17syn$^+$. *J. Virol.*, **64**, pp. 1014–1020.

Chou,J., Kern,E.R., Whitley,R.J., and Roizman,B. (1990) Mapping of herpes simplex virus-1 neurovirulence to γ134.5, a gene nonessential for growth in culture. *Science,* **250**, pp. 1262–1266.

Chou,J., and Roizman,B. (1992) The γ134.5 gene of herpes simplex virus 1 precludes neuroblastoma cells from triggering total shutoff of protein synthesis characteristic of programmed cell death in neuronal cells. *Proc. Natl. Acad. Sci. USA,* **89**, pp. 3266–3270.

Chou,J., Poon,A.P., Johnson,J., and Roizman,B. (1994) Differential response of human cells to deletions and stop codons in the γ134.5 gene of herpes simplex virus. *J. Virol.*, **68**, pp. 8304–8311.

Chou,J., and Roizman,B. (1994) Herpes simplex virus 1 γ134.5 gene function, which blocks the host response to infection, maps in the homologous domain of the genes expressed during growth arrest and DNA damage. *Proc. Natl. Acad. Sci. USA,* **91**, pp. 5247–5251.

Chou,J., Chen,J.J., Gross,M., and Roizman,B.. (1995) Association of a M_r 90,000 phosphoprotein with protein kinase protein kinase R in cells exhibiting enhanced phosphorylation of translation initiation factor eIF-2α and premature shutoff of protein synthesis after infection with γ134.5$^-$ mutants of herpes simplex virus 1. *Proc. Natl. Acad. Sci. USA*, **92**, pp. 10516–10520

Claverie, J.-M. (1996) Effective large-scale sequence similarity searches. *Methods Enzymol*. **266**, 212-226

Claverie,J.M. (1997). Computational methods for the identification of genes in vertebrate genomic sequences. *Hum Mol Genet.*, **6**(10), pp. 1735-1744.

Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S., Barry,C.E., et al. (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, **393,** pp. 537-544

Davenport,D.S., Johnson,D.R., Holmes,G.P, Jewett,D.A., Ross,S.C. and Hilliard,J.K.(1994) Diagnosis and management of human B virus (Herpesvirussimiae) infections in Michigan. *Clin. Infect. Dis.*, **19**, pp. 33–41.

Davison,A.J., Dolan,A., Akter,P., Addison,C., Dargan,D.J., Alcendor,D.J., McGeoch,D.J., and Hayward,G.S. (2003) The human cytomegalovirus genome revisited: comparison with the chimpanzee cytomegalovirus genome. *J. Gen. Virol.*, **84**, pp. 17–28.

Deiss,L.P., Chou,J., and Frenkel,N. (1986). Functional domains within thea sequence involved in the cleavage-packaging of herpes simplex virus DNA. *J. Virol.*, **59**, pp 605–618.

Dolan,A., Jamieson, F.E., Cunningham,C., Barnett,B.C., and Mc-Geoch,D.J. (1998). The genome sequence of herpes simplex virus type 2. *J. Virol.*, **72**, pp. 2010–2021.

Dolyniuk,M., Pritchett,R., and Kieff,E. (1976) Proteins of Epstein-Barr virus. I. Analysis of the polypeptides of purified enveloped Epstein-Barr virus. *J. Virol.*, **17**, pp. 935–949.

Dominguez,G., Dambaugh, T.R., Stamey,F.R., Dewhurst,S., Inoue,N. and Pellett,P.E. (1999) Human herpesvirus 6B genome sequence: coding content and comparison with human herpesvirus 6A. *J. Virol*., **73**, pp 8040–8052.

Dong, S., Searls, D.B. (1994) Gene structure prediction by linguistic methods. *Genomics* **23**, 540-551

Durbin,S., Eddy,A., Krogh,G. and Mitchison,G. (1998) Biological Sequence Analysis. Cambridge University Press, Cambridge, UK.

Duret,L. (2000) tRNA gene number and codon usage in the C. elegans genome are coadapted for optimal translation of highly expressed genes. *Trends Genet.*, **16**, pp. 287–289.

Duret,L., and Mouchiroud,D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. *Natl. Acad. Sci. USA*, **96**, pp. 4482–4487.

Elias,P., Gustafsson,C.M., and Hammarsten,O. (1990) The origin binding protein of herpes simplex virus 1 binds cooperatively to the viral origin of replication oris. *J. Biol. Chem.*, **265**, pp 17167–17173.

Farmer,A.D., Calef,C.E., Millman,K. and Myers,G.L. (1995) The Human Papillomavirus Database. *J. Biomed. Sci.*, **2**, pp. 90-104.

Fickett, J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucl. Acid Res.*, **10**, pp. 5303-5318.

Fickett, J.W. (1995) The gene identification problem: An overview for developers. *Comput. Chem*., **20**, pp. 103-118

Fickett, J. W. (1996) Finding genes by computer: the state of the art. *Trends Genet.*, **12**, 316-320

Fields, C. A., Soderlind, C. A. (1990) GM: a pratical tool for automating DNA sequence analysis. *Comput. Appl. Biosci*. **6**, 263-270

Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M., et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**(5223), pp. 496.512

Fleming,P., Davis-Poynter,N., Degli-Esposti,M., Densley,E., Papadimitriou,J., Shellam,G., and Farrell,H. (1999) The Murine Cytomegalovirus chemokine homolog, m131/129, is a determinant of viral pathogenicity. *J. Virology*, **73(8)**, 6800-6809

Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M., Miller,W., Ford,M.E., Sarkis,G.J. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**(9), pp. 967-74.

Fraser,C.M., Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, et al.(1995) The minimal gene complement of Mycoplasma genitalium. Science, **270**(5235), pp. 397-403

Frishman, D., A. Mironov, H.-W. Mewes, M. Gelfand (1998). "Combining diverse evidence for gene recognition in completely sequenced bacterial genomes". *Nucleic Acids Res*., **26**, 2941-2947

Gelfand, M.S. (1990) Computer prediction of the exon-intron structure of mammalian pre-mRNAs. *Nucl. Acid Res.*, **18**, pp. 5865-5869

Gelfand, M.S. (1995) Prediction of function in DNA sequence analysis. *J. Comp. Biol*., **2**, pp. 87-115.

Gelfand, M.S., Mironov, A.A., Pevzner, P.A. (1996) Gene recognition via spliced sequence alignment. Proc. Natl. Acad. Sci. USA 93, 9061-9066

Gelfand, M.S., Roytberg, M.A. (1993) Prediction of the exon-intron structure by a dynamic programming approach. *BioSystems,* **30**, 173-182.

Gibson,W., Baxter,M.K., and Clopper,K.S. (1996) Cytomegalovirus missing capsid protein identified as heat-agregable product of human cytomegalovirus UL46. *J. Virol.*, **70**, pp. 7454–7461.

Gish, W., States, D. J. (1993) "Identification of protein coding regions by database similarity search." *Nature Genet*. **3**, 266-272
Gouy,M., and Gautier,C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res*., **10**, pp. 7055–7074.

Gotoh, O., (2000). "Homology-based gene structure prediction: simplified matching algorithm using a translated codon (tron) and improved accuracy by allowing for long gaps". *Bioinformatics*, **16**(3):190-202

Gray,C.P., and Kaerner,H.C. (1984) Sequence of the putative origin of replication in the UL region of herpes simplex virus type 1 ANG DNA. *J. Gen. Virol.*, **65,** pp 2109–2119.

Gribskov, M., Devereux, J., Burgess, R. R. (1984) The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. *Nucl. Acid Res*. **12**, 539-549.

Guigo, R., Knudsen, S., Drake, N., Smith, T. (1992) Prediction of gene structure. *J. Mol. Biol*. **225**, 141-157

Guigo,R. (1997). Computational gene identification: an open problem. *Comput Chem.*, **21**(4), pp. 215-222.

Habib,T, Kirby,J, and Sweet,C. (2003) Analysis of murine cytomegalovirus (MCMV) genes. CMV2003 Conference, Maastricht, Netherlands

Hambly,E., Tetart,F., Desplats,C., Wilson,W.H., Krisch,H.M. and Mann,N.H. (2001) A conserved genetic module that encodes the major virion components in both the coliphage T4 and the marine cyanophage S-PM2. *Proc. Natl Acad. Sci. USA*, **98**, pp . 11411-11416.

Hardwicke,M.A., and Schaffer,P.A. (1995) Cloning and characterization of herpes simplex virus type 1 oriL: comparison of replication and protein-DNA complex formation by oriL and oriS. *J. Virol.*, **69**, pp 1377–1388.

Harvey,D.M., and Levine,A.J. (1991) p53 alteration is a common event in the spontaneous immortalization of primary BALB/c murine embryo fibroblasts. *Genes Dev.*, **5**, pp. 2375–2385.

Harrington, L., Wall,L.V., and Kelly,D.C. (1992) Molecular cloning and physical mapping of the genome of simian herpes B virus and comparison of genome organization with that of herpes simplex virus type 1. *J. Gen. Virol.*, **73**, pp 1217–1226.

Hazuda,D.J., Perry,H.C, Naylor,A.M, and McClements,W.L. (1991) Characterization of the herpes simplex virus origin binding protein interaction with OriS. *J. Biol. Chem.*, **266**, pp 24621–24626.

He,B., Chou,J., Brandimarti,R., Mohr,I., Gluzman,Y., and Roizman,B. (1997) Suppression of the phenotype of γ134.5$^-$ herpes simplex virus 1: failure of activated RNA-dependent protein kinase to shut off protein synthesis is associated with a deletion in the domain of the α47 gene. *J. Virol.*, **71**, pp. 6049– 6054.

He,B., Gross,M., and Roizman,B. (1997) The γ134.5 protein of herpes simplex virus 1 complexes with protein phosphatase 1α to dephosphorylate the alpha subunit of the eukaryotic translation initiation factor 2 and preclude the shutoff of protein synthesis by double-stranded RNA-activated protein kinase. *Proc. Natl. Acad. Sci. USA*, **94**, pp. 843–848.

He,J.G., L,L., Deng,M., He,H.H., Weng,S.P., Wang,X.H., Zhou,S.Y., Long,Q.Z., Wang,X.Z. and Chan,S.M. (2002) Sequence analysis of the complete genome of an iridovirus isolated from the tiger frog. *Virology*, **292**, pp. 185-197.

Hein,J. (1990) Unified approach to alignment and phylogenies. *Methods Enzymol.*, **183**, pp 626–645.

Henderson J., S. Salzberg, K. H. Fasman (1997). "Finding genes in DNA with a Hidden Markov Model.". *J Comput Biol*., **4**(2):127-141.

Hiscock,D. and Upton,C. (2000) Viral Genome DataBase: storing and analyzing genes and proteins from complete viral genomes. *Bioinformatics*, **16**, pp. 484-485.

Huang X., M. D. Adams, H. Zhou, A. R. Kerlavage (1997). "A tool for analyzing and annotating genomic Sequences". *Genomics*, **46**(1): 37-45.

Huber,M.T., and Compton,T. (1999) Intracellular formation and processing of the heterotrimeric gH-gL-gO (gCIII) glycoprotein envelope complex of human cytomegalovirus. *J. Virol.*, **73**, pp 3886–3892.

Hutchinson, G. B., Hayden, M. R. (1992) The prediction of exons through an analysis of spliceable open reading frames. Nucl. Acid Res. **20**, 3453-3462

Ikemura,T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.,* **2**, pp. 13–34.

Jiang J., H. J. Jacob (1998). "EbEST: an automated tool using expressed sequence tags to delineate gene structure". *Genome Res*., **8**(3):268-275.

Kan, Z., E.C. Rouchka, W.R. Gish, D.J. States (2001). "Gene structure prediction and alternative splicing analysis using genomically aligned ESTs". *Genome Res*., **11**(5):889-900.

Kattenhorn,L.M., Mills,R., Wagner,M., Lomsadze,A., Makeev,V., Borodovsky,M., Ploegh,H.L., Kessler,B.M. (2004) Identification of proteins associated with murine cytomegalovirus virions. *J. Virol.*, **78** (20), pp. 11187-11197

Keeble, S.A., Christofinis,G.J., and Wood,W. (1958). Natural B virus infection in rhesus monkeys. *J. Pathol. Bacteriol.*, **76**, pp 189–199.

Kieff,E. and Rickinson,A.B. (2001) Epstein±Barr virus and its replication. In Knipe,D.M. and Howley,P.M. (eds), Fields Virology. Lippincott Williams and Wilkins, Philadelphia, PA, Vol. 2, pp. 2511-2672.

Killeen,A.M., Harrington,L., Wall,L.V., and Kelly,D.C.. (1992). Nucleotide sequence analysis of a homologue of herpes simplex virus type 1 gene US9 found in the genome of simian herpes B virus. *J. Gen. Virol.*, **73**, pp 195–199.

Kingham, B.F., Zelnik,V, Kopacek,J., Majerciak,V., Ney,E., and Schmidt,C.J. (2001) The genome of herpesvirus of turkeys: comparative analysis with Marek's disease viruses. *J. Gen. Virol*, **82**, pp 1123–1135.

Kinter,M., and Sherman,N.E. (2000) Protein sequencing and identification using tandem mass spectrometry. Wiley & Sons, New York, N.Y.

Kornberg,T.B. and Krasnow,M.A. (2000) The Drosophila genome sequence: implications for biology and medicine. *Science*, **287**, pp. 2218-20.

Krogh A., M. Brown, I. S. Mian, K. Sjolander, D. Haussler (1994). "Hidden Markov models in computational biology. Applications to protein modeling". *J Mol Biol*., **235**(5): 1501-1531.

Krogh, A. (1997) "Two methods for improving preformance of an HMM and their application for gene finding". In: *Proc. Fifth Int. Conf. Intelligent Systems for molecular Biology*. Gaasterland T. et al. Eds. AAAI Press p179-186

Kulp D., D. Haussler, M. G. Reese, F. H. Eeckman (1996). "A generalized hidden Markov model for the recognition of human genes in DNA". *Proc Int Conf Intell Syst Mol Biol*., **4**: 134-142.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., Funke,R., Gage,D., Harris,K., et al. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), pp. 860-921

Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, pp. 208-214.

Leib,D.A., Machalek,M.A., Williams,B.R.G., Silverman,R.H., and Virgin,H.W. (2000) From the cover: specific phenotypic restoration of an attenuated attenuated virus by knockout of a host resistance gene. *Proc. Natl. Acad. Sci. USA*, **97**, pp. 6097–6101.

Ludwig, H., Pauli,G., Gelderblom,H.R., Darai,G., Koch,H-G., Flugel,R.M., Norrild,B., and Daniel,M.D. (1983) B virus (Herpesvirus simiae) In B. Roizman (ed.), The herpesviruses, vol. 2. Plenum Press, New York, N.Y., pp. 385–428.

Lukashin,A.V., and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, **26**, pp1107–1115.

Mar,E.C., Patel,P.C., and Huang,E.S. (1981) Human cytomegalovirusassociated DNA polymerase and protein kinase activities. *J. Gen. Virol.*, **57**, pp 149–156.

Mar Alba,Á,M., Lee,D., Pearl,F.M.G., Shepherd,A.J., Martin,N., Orengo,C.A. and Kellam,P. (2001) VIDA: a virus database system for the organisation of virus genome open reading frames. *Nucleic Acids Res.*, **29**, pp. 133-136.

Majoros William H, Mihaela Pertea, Corina Antonescu, Steven L Salzberg (2003). "GlimmerM, Exonomy and Unveil: three ab initio eukaryotic genefinders". *Nucleic Acids Res*., **31**:3601-3604.

Majoros,W.H., Pertea,M., Delcher,A.L., and Salzberg,S.L. (2005) Efficient decoding algorithms for generalized hidden Markov model gene finders. *BMC Bioinformatics*, **6**, pp. 15

Martin,D.W., Deb,S.P., Klauer,J.S., and Deb,S. (1991) Analysis of the herpes simplex virus type 1 OriS sequence: mapping of functional domains. *J. Virol.*, **65**, pp 4359–4369.

Mathe,C., Sagot,M.F., Schiex,T., Rouze,P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.*, **30**, pp. 4103-4117.

McGeoch,D.J., Cunningham,C., McIntyre,G., and Dolan,A.. (1991) Comparative sequence analysis of the long repeat regions and adjoining parts of the long unique regions in the genomes of herpes simplex viruses types 1 and 2. *J. Gen. Virol.*, **72**, pp 3057–3075.

McGeoch,D.J., Dolan,A., Donald,S., and Brauer,D.H. (1986) Complete DNA sequence of the short repeat region in the genome of herpes simplex virus type 1. *Nucleic Acids Res.*, **14**, pp 1727–1745.
Keeble, S. A. (1960) B virus infection in monkeys. *Ann. N.Y. Acad. Sci.*, **85**, pp 960–969.

McVoy,M.A., Nixon,D.E, Adler,S.P, and Mocarski,E.S. (1998) Sequences within the herpesvirus-conserved pac1 and pac2 motifs are required for cleavage and packaging of the murine cytomegalovirus genome. *J. Virol.*, **72**, pp 48–56.

Mehdi,H., Ono,E., and Gupta,K.C. (1990) Initiation of translation at CUG, GUG, and ACG codons in mammalian cells. *Gene*, **91**, pp 173–178.

Mesyanzhinov,V.V., Robben,J., Grymonprez,B., Kostyuchenko,V.A., Bourkaltseva,M.V., Sykilinda,N.N, Krylov,V.N., and Volckaert,G. (2002) The genome of bacteriophage phiKZ of Pseudomonas aeruginosa. *J. Mol. Biol.*, **317**, pp. 1–19.

Milanesi,L., Kochanov,N.A., Rogozin,I B., Ischenko,I.V., Kel,A.E., Orlov,Y L., Ponomarenko,M.P., Vezzoni,P. (1993) Gen Viewer: a computing tool for protein-coding regions prediction in nucleotide sequences. In: Proceedings of the Second International

Conference on Bioinformatics, Supercomputing and Complex Genome Analysis. Lim, H. A. et al. Eds. World Scientific, Singapore, pp. 573-588

Milanesi,L., Rogozin,I.B. (1998) Prediction of human gene structure, In: Guide to Human Genome Computing, 2nd edn. Bishop, M. J. Ed. Academic Press, Cambridge, pp. 215-259

Mills,R., Rozanov,M., Lomsadze,A., Tatusova,T., and Borodovsky,M. (2003) Improving gene annotation of complete viral genomes. *Nucleic Acids Res.*, **31**, pp. 7041–7055.

Mironov, A. A., Roytberg, M. A., Pevzner, P. A., Gelfand, M. (1998) Performance-guarantee gene predictions via spliced alignment. *Genomics*. **51**: 332-339

Mocarski,E.S., and Courcelle,C.T. (2001) Cytomegaloviruses and their replication, p. 2629–2673. In Knipe,D.M., Howley,P.M., Griffin,D.E., Lamb,R.A., Martin,M.A., Roizman,B., and Straus,S.E. (ed.), Fields virology, vol. 2. Lippincott-Raven, Philadelphia, Pa.

Mohr,I., and Gluzman,Y. (1996) A herpesvirus genetic element which affects translation in the absence of the viral GADD34 function. *EMBO J.*, **15**, pp. 4759–4766.

Moriyama,E. N., and Powell,J.R. (1997) Codon usage bias and tRNA abundance in Drosophila. *J. Mol. Evol.*, **45**, pp. 514–523.

Mott R. (1997). "EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA". *Comput Appl Biosci*., **13**(4):477-478.

Murphy,E., Rigoutsos,I., Shibuya,T., and Shenk,T.E. (2003) Reevaluation of human cytomegalovirus coding potential. *Proc. Natl. Acad. Sci. USA*, **100**, pp. 13585–13590.

Nealon,K., Newcomb,W.W., Pray,T.R., Craik,C.S., Brown,J.C., and Kedes,D.H. (2001) Lytic replication of Kaposi's sarcoma-associated herpesvirus results in the formation of multiple capsid species: isolation and molecular characterization of A, B, and C capsids from a gammaherpesvirus. *J. Virol.*, **75**, pp.2866–2878.

Ohsawa,K., Black,D.H., Sato,H. and Eberle,R. (2002) Sequence and genetic arrangement of the US region of the monkey B virus (cercopithecine herpesvirus 1) genome and comparison with the US regions of other primate herpesviruses. *J. Virol.*, **76**, pp.1516–1520.

Oliveira,S.A., and Shenk,T.E.(2001) Murine cytomegalovirus M78 protein, a G protein-coupled receptor homologue, is a constituent of the virion and facilitates accumulation of immediate-early viral mRNA. *Proc. Natl. Acad. Sci.* USA, **98**, pp. 3237–3242.

Olivo,P.D., Nelson,N.J., and Challberg,M.D. (1988) Herpes simplex virus DNA replication: the UL9 gene encodes an origin-binding protein. *Proc. Natl. Acad. Sci. USA*, **85**, pp 5414–5418.

Palmer,A.E. (1987) B virus, herpesvirus simiae: historical perspective. *J. Med. Primatol.*, **16**, pp 99–130.

Pari,G.S., and Anders,D.G. (1993) Eleven loci encoding trans-acting factors are required for transient complementation of human cytomegalovirus ori-Lyt-dependent DNA replication. *J. Virol.*, **67**, pp. 6979–6988.

Parra, G., E. Blanco, R. Guigo (2000). "GeneID in Drosophila". *Genome Res*., **10**(4):511-515.

Patrone,M., Percivalle,E., Secchi,M., Fiorina,L., Pedrali-Noy,G., Zoppe,M., Baldanti,F., Hahn,G., Koszinowski,U.H., Milanesi,G., and Gallina,A. (2003) The human cytomegalovirus UL45 gene product is a late, virion-associated protein and influences virus growth at low multiplicities of infection. *J. Gen. Virol.*, **84**, pp. 3359–3370.

Perelygina,L., Zhu,L., Zurkuhlen,H., Mills,R., Borodovsky,M., and Hilliard,J.K. (2003) Complete sequence and comparative analysis of the genome of herpes B virus (Cercopithecine herpesvirus 1) from a rhesus monkey. *J. Virol.*, **77**(11), pp. 6167-77

Perelygina,L., Patrusheva,I., Manes,N., Wildes,M.J., Krug,P., and Hilliard,J.K. (2003) Quantitative real-time PCR for detection of monkey B virus (cercopithecine herpesvirus 1) in clinical samples. *J. Virol. Methods*, **109**, pp 245–251.

Perkins,D.N., Pappin,D.J., Creasy,D.M., and Cottrell,J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, pp. 3551–3567.

Pietropaolo,R.L., and Compton,T. (1997) Direct interaction between human cytomegalovirus glycoprotein B and cellular annexin II. *J. Virol.*, **71**, pp. 9803–9807.

Ramensky,V.E., Makeev,V., Roytberg,M.A., and Tumanyan,V.G. (2000) DNA segmentation through the Bayesian approach. *J. Comput. Biol.*, **7**, pp. 215–231.

Randall,G., Lagunoff,M., and Roizman,B. (1997) The product of ORF O located within the domain of herpes simplex virus 1 genome transcribed during latent infection binds to and inhibits in vitro binding of infected cell protein 4 to its cognate DNA site. *Proc. Natl. Acad. Sci. USA*, **94**, pp 10379–10384.

Randall,G., and Roizman,B. (1997) Transcription of the derepressed open reading frame P of herpes simplex virus 1 precludes the expression of the antisense 134.5 gene and may account for the attenuation of the mutant virus. *J. Virol.*, **71**, pp 7750–7757.

Rawlinson,W.D., Farrell,H.E., and Barrell,B.H. (1996) Analysis of the complete DNA sequence of murine cytomegalovirus. *J. Virol.*, **70**, pp. 8833–8849.

Recketenwald,J. and Schmidt,H. (2002) The nucleotide sequence of Shiga toxin (Stx) 2e-encoding phage FP27 is not related to other Stx phage genomes, but the modular genetic structure is conserved. *Infect. Immun.*, **70**, pp. 1896-1908.

Reese, M. G., G. Hartzell, N. L. Harris, U. Ohler, J. F. Abril, S. E. Lewis (2000). "Genome annotation assessment in Drosophila Melanogaster". *Genome Res*., **10**(4): 483-501.

Resch,G., KulikE.M., Dietrich,F.S., and Meyer, E. (1994) Complete genomic nucleotide sequence of the temperate bacteriophage Aa phi 23 of Actinobacillus actinomycetemcomitans. *J. Bacteriol*, **186** (16), pp. 5523-5528

Robison, K., Gilbert, W., Church, G. M. (1994) "Large scale bacterial gene discovery by similarity search". Nature Genetics **7**, 205-214.
Roizman,B., and Knipe,D.M. (2001) Herpes simplex viruses and their replication. In D. M. Knipe and P. M. Howley (ed.), Fields virology, 4th ed. Lippincott-Raven Publishers, Philadelphia, Pa., pp. 2399–2460.

Rogozin I. B., L. Milanesi, N. A. Kolchanov (1996). "Gene structure prediction using information on homologous protein sequence". *Comput Appl Biosci*., **12**(3):161-170.

Roizman,B., and Pellett,P.E. (2001) The family Herpesviridae: a brief introduction. In D. M. Knipe and P. M. Howley (ed.), Fields virology. Lippincott-Raven Publishers, Philadelphia, PA , pp 2381–2397.

Salamov, A.A., V.V. Solovyev (2000). "Ab initio gene finding in Drosophila genomic DNA". *Genome Res*., **10**(4);516-522.

Salzberg S.L., Delcher A.L., Kasif S., White O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*. **26**, 544-548.

Salzberg, S., A. L. Delcher, K. H. Fasman, J. Henderson (1998). "A decision tree system for finding genes in DNA". *J Comput Biol*., **5**(4): 667-680

Sambrook,J., Fritsch,E.F and Maniatis,T. (1989) Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Sanger,F., Air,G.M., Barrell,B.G., Brown,N.L., Coulson,A.R., Fiddes,C.A., Hutchison,C.A., Slocombe,P.M., and Smith,M. (1977) Nucleotide sequence of bacteriophagephi X174 DNA. *Nature*, **265**(5596), pp. 687-95

Schlueter, S. D., Q. Dong, V. Brendel (2003). "GeneSeqer@PlantGDB: gene structure prediction in plant genomes". *Nucleic Acids Res*., **31**:3597-3600

Sharp,P.M., Tuohy,T.M., and Mosurski,K.R. (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.*, **14**, pp. 5125–5143.

Shepherd, J. C. W. (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justifications. *Proc. Natl. Acad. Sci. USA,***78**, 1596-1600.

Singh,R., Haghjoo,E., and Liu,F. (2003) Cytomegalovirus M43 gene modulates T helper cell response. *Immunol. Lett.*, **88**, pp. 31–35.

Slomka, M.J., Harrington,L., Arnold,C., Norcott,J.P. and Brown,D.W. (1995) Complete nucleotide sequence of the herpesvirus simiae glycoprotein G gene and its expression as an immunogenic fusion protein in bacteria. *J. Gen. Virol.*, **76**, pp 2161–2168.

Smith, A.L., Black,D.H., and Eberle,R. (1998) Molecular evidence for distinct genotypes of monkey B virus (herpesvirus simiae) which are related to the macaque host species. *J. Virol.*, **72**, pp 9224–9232.

Snyder, E.E., Stormo, G.D. (1995a) Identifying genes in genomic DNA sequences. In: Nucleic Acid and Protein Sequence Analysis: A Pratical Approach, 2nd edn. IRL Press, Oxford

Snyder, E.E. and G.D. Stormo (1995b). Identification of protein coding regions in genomic DNA. *J Mol Biol.*, **248**(1), pp. 1-18.

Solovyev, V. V., Salamov, A. A., Lawrence, C. B. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucl. Acid Res*. **22**, 5156-5163

Staden, R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucl Acid Res*. **12**, 505-519.

Stanke, M., S. Waack (2003). "Gene prediction with a hidden Markov model and a new intron submodel". *Bioinformatics*, **19**:ii215-ii225.

States, D. J., Gish, W. (1994) "Combined use of sequence similarity and codon bias for coding region identification". *J. Comp. Biol*. **1**, 39-50.

Stormo,G.D. (2000). Gene-finding approaches for eukaryotes. *Genome Res.*, **10**(4), pp. 394-397.

Taha,M.Y., Clements,G.B., and Brown,S.M. (1989) A variant of herpes simplex virus type 2 strain HG52 with a 1.5 kb deletion in RL between 0 to 0.02 and 0.81 to 0.83 map units is non-neurovirulent for mice. *J. Gen. Virol.*, **70**, pp. 705–716.

Takai,S., Hines,S.A., Sekizaki,T., Nicholson,V.M., Alperin,D.A., Osaki,M., Takamatsu,D., Nakamura,M., Suzuki,K., Ogino,N., Kakuda,T., Dan,H., and Prescott,J.F. (2000) DNA sequence and comparison of virulence plasmids from Rhodococcus equi ATCC 33701 and 103. *Infect. Immun.*, **68**, pp. 6840–6847.

Thomas, A., Skolnick, M. H. (1994) A probabilistic model for detecting coding regions in DNA sequences. IMA *J. of Math. Applied in Med. & Biol*. **11**, 149-160.

Thompson,R. L., Rogers,S.K., and Zerhusen,M.A. (1989) Herpes simplex virus neurovirulence and productive infection of neural cells is associated with a function which maps between 0.82 and 0.832 map units on the HSV genome. *Virology,* **172**, pp. 435–450.

Tomb,J-F, White,O., Kerlavage,A.R., et al. (1997) The complete sequence of the gastric pathogen Helicobacter pylori. *Nature*, **388**, pp. 539-547

Tu,A.H., Voelker,L.L., Shen,X., and Dybvig,K. (2001) Complete nucleotide sequence of the mycoplasma virus P1 genome. *Plasmid*, **45**, pp. 122–126.

Tulman,E.R., Afonso,C.L., Lu,Z., Zsak,L., Rock,D.L., and Kutish,G.F. (2000) The genome of a very virulent Marek's disease virus. *J. Virol*, **74**, pp 7980–7988.

Uberbacher, E. C., Mural, R. J. (1991) A probabilistic model for detecting coding regions in DNA sequences. IMA *J. Math. Appl. Med. Biol*. **11**, 149-160

Usuka, J., W. Zhu, V. Brendel (2000). "Optimal spliced alignment of homologous cDNA to a genomic DNA template". *Bioinformatics*, **16**(3):203-211.

Varmuza,S.L., and Smiley,J.R. (1985) Signals for site-specific cleavage of HSV DNA: maturation involves two separate cleavage events at sites distal to the recognition sequences. *Cell*, **41,** pp 793–802.

Venkatesan,M.M., Goldberg,M.B., Rose,D.J., Grotbeck,E.J., Burland,V. and Blattner,F.R. (2001) Complete DNA sequence and analysis of the large virulence plasmid of Shigella felxneri. *Infect. Immun.*, **69**, pp. 3271-3285.

Ventor,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A., Gocayne,J.D., et al. (2001) The sequence of the human genome. *Science*, **291**(5507), pp. 1304-51

Vink,C., Beuken,E., and Bruggeman,C.A. (2000) Complete DNA sequence of the rat cytomegalovirus genome. *J. Virol.*, **74**, pp 7656–7665.

Virgin,H., Latreille,P., Wamsley,P., Hallsworth,K., Weck,K.E., Dal Canto,A.J., and Speck,S.H. (1997) Complete sequence and genomic analysis of murine gammaherpesvirus 68. *J. Virol.*, **71**, pp 5894–5904.

Volfovsky, N., B.J. Haas, S.L. Salzberg (2003). "Computational discovery of internal micro-exons". *Genome Res*., **13**(6):1216-1221.

Wagner,M., Jonjic,S., Koszinowski,U.H., and Messerle,M. (1999) Systematic excision of vector sequences from the BAC-cloned herpesvirus genome during virus reconstitution. *J. Virol.*, **73**, pp. 7056–7060.

Wagner,M., Gutermann,A., Podlech,J., Reddehase,M.J., and Koszinowski,U.H. (2002) Major histocompatibility complex class I allele-specific cooperative and competitive interactions between immune evasion proteins of cytomegalovirus. *J. Exp. Med.*, **196**, pp. 805–816.

Walboomers,J.M., and Schegget,J.T. (1976) A new method for the isolation of herpes simplex virus type 2 DNA. *Virology*, **74**, pp 256–258.

Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P., Antonarakis,S.E., Attwood,J., Baertsch,R. et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**(6915), pp. 520-62

Weigler, B.J. (1992) Biology of B virus in macaque and human hosts: a review. *Clin. Infect. Dis.*, **14**, pp 555–567.

Wheelan S. J., D. M. Church, J. M. Ostell (2001). "Spidey: a tool for mRNA-to-genomic alignments". *Genome Res*., **11**(11):1952-1957.

Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova,T.A., Wagner,L. and Rapp,B.A. (2001) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **29**, pp .11-16.

Whitley,R.J., and Hilliard,J.K. (2001) Cercopithecine herpesvirus (B virus), In D. M. Knipe and P. M. Howley (ed.), Fields virology. Lippincott-Raven Publishers, Philadelphia, Pa. pp. 2835–2848.

Wu,C.A., Carlson,M.E., Henry,S.C., and Shanley,J.D. (1999) The murine cytomegalovirus M25 open reading frame encodes a component of the tegument. *Virology*, **262**, pp. 265–276.

Xu, Y., R. Mural, M. Shah, E. Uberbacher (1994). "Recognizing exons in genomic sequence using GRAIL II". *Genet Eng (N Y)* **16**: 241-253.

Xu,J., Scalzo,A.A., Lyons,P.A., Farrell,H.E., Rawlinson,W.D., and Shellam,G.R. (1994) Identification, sequencing and expression of the glycoprotein L gene of murine cytomegalovirus. *J. Gen. Virol.*, **75**, pp. 3235–3240.

Yang,F., He,J., Lin,X., Li,Q., Pan,D., Zhang,X. and Xu,X. (2001) Complete genome sequence of the shrimp white spot bacilliform virus. *J. Virol.*, **75**, pp. 11811-11820.

Zhan,X., Lee,M., Abenes,G., Von Reis,I., Kittinunvorakoon,C., Ross-Macdonald,P., Snyder,M., and Liu,F. (2000) Mutagenesis of murine cytomegalovirus using a Tn3-based transposon. *Virology*, **266**, pp 264–274.

Zhang, M. Q. (1997) "Identification of protein coding regions in the human genome by quadratic discriminant analysis". *Proc Natl Acad Sci USA*., **94**(2): 565-568.

Zhang,M.Q. (2002) Computational prediction of eukaryotic protein-coding genes. *Nat Rev Genet.*, **3**(9), pp. 698-709.

Zimmermann,W., Broll,H., Ehlers,B., Buhk,H.J., Rosenthal,A., and Goltz,M. (2001) Genome sequence of bovine herpesvirus 4, a bovine rhadinovirus, and identification of an origin of DNA replication. *J. Virol.*, **75**, pp. 1186–1194.